"The fact that the majority seems to be…"

A corpus-driven investigation of lexical bundles in native and non-native academic English

Jonas Lie

A Thesis Presented to the Department of Literature, Area Studies and European Languages

UNIVERSITY OF OSLO
Supervisor: Professor Hilde Hasselgård

in Partial Fulfilment of the Requirements for the MA Degree

December 2013

“The fact that the majority seems to be…”

A corpus-driven investigation of lexical bundles in native and non-native English

Jonas Lie

2013

”The fact that the majority seems to be… - A corpus-driven investigation of lexical bundles in native and non-native academic English”

Jonas A. Lie

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Acknowledgements

I would like to give my heartfelt thanks to Professor Hilde Hasselgård for all her much appreciated and indispensable guidance, ideas, merciless attention to detail and humour in the process of writing this thesis; to all the regulars of the ILOS students' break room, without whom the late nights, early mornings and protracted lunches would have been significantly less enjoyable; to the Bouldering Bros who kept my mind and body limber; and to Andrea for her tireless encouragement, feedback and mind-reading.

Table of Contents

Figures:

Tables:

1 Introduction

The present study aims to investigate the ways in which Norwegian university students use English when writing academic texts, by conducting an in-depth examination of the Norwegian section of the Varieties of English for Specific Purposes dAtabase (VESPA).

The choice of demographic is not merely opportunistic, but chosen specifically because the investigation of the language used by university students offers insights not only into the language of universities, academia and higher education as a whole, but also a unique look at the results of the Norwegian school system's formal instruction in English, from primary school to upper secondary school.

The fundamental premise of the study is, of course, one of exquisite irony: an academic text investigating how writers with English as their second language use English in academic texts, all by a writer with English as his second language. I am, however, willing to embrace this, and accept the crippling shame that might follow from my own failure to adhere to the standards I prescribe, because it allows me to use my academic infatuation with the fields of corpus linguistics and phraseology to the benefit of my future career in language teaching.

This fascination with corpus linguistics is owed largely to the unique insights a corpus offers into the staggering diversity and mutability of language, and the way in which it illustrates this through *actual* language produced by *actual* people in *actual* situations, far removed from the stringent rules, conventions, order and logic of prescriptive linguistics.

A natural companion to corpus linguistics, phraseology seeks, by investigating how words behave around other words, to answer the classic conundrum of why something composed from perfectly good words and ordered into a perfectly acceptable sentence can still seem so fundamentally alien to a native speaker; it seeks to provide an answer to the seemingly innocent question that haunts the dreams of any language teacher, the question that so often follows having corrected a student because a construction seemed slightly "off": "Why?"

This study is inspired in particular by the work of Douglas Biber, and specifically his work – in a variety of collegial constellations – with *lexical bundles*. Lexical bundles, are, in

Biber's own terms, simply "the most frequent recurring fixed lexical sequences in a register" (Biber & Conrad 2004:59). By examining these, we can discover what constructions form the bricks and mortar of academic language, and hopefully gain from these some insights into how learners can improve their academic English, and even more importantly: How we as teachers can help them to do so.

This study will consist, in addition to a presentation of the relevant theories and material, of two parts: The first is a general overview of the two corpora, followed by a select few in-depth investigations of items that stand out as particularly phraseologically interesting. In doing so, I hope to be able to answer three questions:

1) To what extent do the VESPA contributors use lexical bundles?

2) What functions do these lexical bundles serve, and how does their use and distribution compare to that of native speakers?

3) In what areas are VESPA contributors over- under or misusing lexical bundles? Do any patterns emerge?

2 Theory

2.1 Corpus Linguistics

"A helluva lot of words, stored on a computer" - Geoffrey Leech (1992:106)

Although perhaps somewhat lacking in terms of precision and the expected academic finesse, Leech's humorous one-liner quite aptly captures the essence of corpus linguistics: The collection of a body of text - the *corpus* - to which the linguist can apply the tools of her trade in order to illuminate some aspect of language, be it technical, stylistic or artistic. Armed with a corpus, the linguist is not only able to distill from it an idea or theory of language, but can also refer to situations in which language behaves in accordance with her claims. Regardless of whether one sees corpus linguistics as a method or a theory in itself - a distinction that will be discussed more in depth below - this reliance on *attested* language is what sets corpus linguistics apart. It is the study not of how a language *can* or *should* be used, but the study of how language *is* used.

Despite being relatively recently automated and applied to linguistics, the basic principle of corpus linguistics – the systematic collection and contextual reorganisation of texts - is centuries old, finding its genesis among thirteenth-century Dominican monks, who compiled the Bible's wealth of references to places, prophets, genealogies and theological concepts into complex biblical *concordances* that would aid in their exegeses. These early concordances mapped every single occurrence of a word throughout the Bible, and are as such similar in form to the earliest linguistic applications of the corpus method, such as the tremendous undertaking of German philologist Käding, who already in 1897 manually analysed a corpus of more than 11 million words. Similar cataloguing methods were frequently employed in early language acquisition studies and pedagogy, where the use of parental diaries detailing children’s speech was the prevalent methodology in the field between 1876-1926 and remain a significant source of normative data even today (McEnery and Wilson 2001:3). In the field of traditional grammar, Danish philologist Otto Jespersen’s seven-volume “A Modern English Grammar on Historical Principles” (1909-49) provided quotes gathered by Jespersen throughout (Haislund 1943), effectively making it one of the

first grammars to rely on attested language.

Beyond the mere mechanics of cataloguing, however, there is little common ground between these early proponents of corpora and the field of modern corpus linguistics. The majority of early corpus users were structuralists, with methodologies firmly anchored to the behaviourist-positivist attitudes that permeated scientific circles until the late 1950's. At the heart of their approach to corpus linguistics was the idea that since all language is governed by a strict set of structural rules, the number of sentences in a language is finite, making it the purpose of a corpus to “collect” these (McEnery & Wilson 2001:7). In broad strokes, they “regarded the corpus as the primary explicandum of linguists” (Leech 91:8), and thus implicitly claimed that a corpus could contain all of language, a notion that inspired vociferous resistance among many contemporaries, most famously by Noam Chomsky, who was thoroughly unimpressed by the supposed merit of corpora:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.
> (Chomsky 1962 in McEnery & Wilson 2001:8)

The debate that followed left corpus linguistics effectively dead in the water, and Chomsky's view on the scientist's intuition as the best means of linguistic investigation became the dominant. (Leech 1991:10)

Despite the fact that his initial criticisms were aimed at notions of language very few linguists would agree with today the falling-out between corpus linguists and those that followed Chomsky in his rejection of corpus evidence as examples of *performance* and not *competence,* and thus of no interest to linguists, has proven schismatic in retrospect, with sides devoting page after page to the still ongoing debate.

The development of and subsequent tremendous advances in both computers and computerised corpora allowed linguists to automate the concordancing process, drastically reducing the time and resources required, enabling them to grapple with more complex problems and larger corpora. Early efforts, such as Kucera and Francis' "Computational Analysis of Present-Day American English" (1967) based on a million-word corpus of texts published in 1961 compiled at Brown University, Rhode Island, demonstrated the flexibility of the Brown electronic corpus, the influence of which can hardly be overstated (Johansson

2008:35). Further work in the field was characterised by extensive international collaboration: Despite English dominating as the subject matter, important contributions have often come from countries where English is a foreign language, with notable contributions from Belgium, the Netherlands, Norway and Sweden. (McEnery & Hardie, 2012:72). Funding these early efforts was tough, however, and corpus work could be a rather nomadic affair, with work starting at one university, but moving on to another once resources are depleted, as was the case with the Lancaster-Oslo-Bergen corpus, another hallmark corpus, designed as a British mirroring of the Brown corpus (Johansson 2008:38).

From the late seventies on, corpus linguistics gained ground exponentially: In the period between 1975 and 1990, the number of articles published more than doubled every five years, with 30 articles published in 1971-1975 against an astounding 320 articles between 1986 and 1990, likely largely owed to the similarly explosive increase in computer processing power. (Johansson 2008:47) Corpus methods were finding their place in many disciplines, and many of the largest contributions to the field of English linguistics from this period have at least one foot in the corpus linguistics camp, such as the corpus-informed landmark grammar of Quirk et al. (1984) and the ground-breaking Collins COBUILD dictionary (1987), a wholly corpus-based project led by John Sinclair, finding both its definitions and examples through corpus work, requiring a far more extensive use of corpora than had previously been possible (Johansson 2008:48).

In the wake of the COBUILD project, Sinclair was also instrumental in another central development in corpus linguistics: The advent of *corpus-driven* studies. With each successive study, and the increasingly complex and diverse corpora that were compiled, there were those that “[...] realised how fundamentally the traditional pre-corpus descriptions of language were being implicitly questioned by the evidence of larger corpora” (Tognini-Bonelli 2001, 85) With access to more and more language in use, researchers realized that *pre-corpus* theories, developed more or less from the reflection and introspection of their authors, were not necessarily wrong, but unable to describe the myriad nuances of how language is used. The problem, in Sinclair's view, was that the methodologies applied to corpora were being used overwhelmingly to test or exemplify pre-existing theories, and that these theories were standing in the way of discovering anything *new* about language: In the same way a fisherman fishing with techniques and equipment developed specifically for salmon in a river famed for its salmon will almost invariably catch salmon, no matter his skill or expertise, a researcher

applying pre-corpus models of language to a corpus will gain little else than pre-corpus insights. It does not seem to be Sinclair's intention to discount those dabbling in linguistic salmon-angling, however, but merely to suggest that throwing a stick of dynamite in said river *could* very well lead to the discovery of some yet to be tasted pelagic treat. His suggested tool had little of the destructive power of dynamite, but he believed it to offer a similar capacity for discovery: Instead of bringing one's conceptions of language to a corpus in order to prove, exemplify or embellish upon some theory, one should investigate the corpus independently, without applying pre-conceived categories. In effect, for those working within the corpus-driven approach, corpus linguistics is not only a method, but a theory in itself, relying on nothing but the corpus or texts themselves, a view neatly summed up in Sinclair's frequently quoted mantra “Trust the text” (Sinclair 2004:23)

A corpus-driven approach is thus hugely reliant on the fidelity of the composition of the corpus at hand, however: if the corpus fails to represent the language it claims to represent in an accurate and balanced manner, the conclusions drawn on the basis of it will be effectively worthless: “The results are only as good as the corpus”, as Sinclair himself puts it (1991:13). The researcher must also know his limits: To approach a corpus entirely without preconception is impossible, as Sinclair was well aware, often drawing on the ideas of J. R. Firth in explaining that “Each scholar makes his own selection and grouping of facts determined by his attitudes and theories and by the nature of his experience of reality of which he himself is part” (J.R. Firth, in Tognini-Bonelli 2001:85). In the community surrounding Sinclair, including such researchers as Michael Hoey, Susan Hunston, Michael Stubbs, Wolfgang Teubert and Elena Tognini-Bonelli, most of whom at some point linked to the University of Birmingham (McEnery & Hardie 2012:122), Firth was a great source of inspiration, enough to earn them the collective label “Neo-Firthians”.

2.2 Learner Corpora

Learner corpus linguistics is concerned primarily with *non-native* language production. “Learners” in this context, and for the purpose of this paper, are both those who learn English in a non-English environment - English as a Foreign Language (EFL) - and those who learn it in a country where English is an official language, regardless of its being a majority language

or not - English as a Second Language (ESL). Their language is unique, at least for the vast majority of learners, as it represents an imperfect recreation of the target language (L2), essentially the learners' approximation of the L2, coloured by aspects their native language (L1). This *interlanguage*, as linguist Larry Selinker termed it (Hasselgård & Johansson 2011:35), constitutes the chief object of study for learner corpus linguistics. Interlanguage is as multifaceted and diverse as the learners who use it, but as these corpora grow, compiling data from hundreds, or even thousands of learners, patterns will eventually start to emerge, and it is the linguist's mission to identify these patterns in order to better understand the complex workings of learner language.

Sylviane Granger (2013:1) suggests four criteria that a dataset should meet in order to qualify as a learner corpus, the first two of which concern its design and the final two its content. In terms of design, the data must first and foremost be in electronic format, disqualifying for example the aforementioned proto-corpora used in early acquisition studies. Secondly, they must be subject to rigorous data collection regimes, since, as Granger puts it, "learner language is obviously highly heterogenous: there are many types of learners and learning situations, and "mixed bag" collections of L2 data present little interest". Such design issues are of course important for any corpus, but becomes doubly so for these corpora, since if we are to make any use of our findings, we must know exactly what variables have affected the interlanguage under investigation. Developing such *learner profiles* can be challenging, however, as there is no universal, agreed-upon measure of proficiency, and as such, indicating the level of a learner is a definite challenge. Going by a system of teacher-graded texts would perhaps seem natural, since such tests are a mainstay of language teaching, but such grading has been found to be highly subjective, with inter-rater agreement often proving unacceptably low (Pendar & Chapelle 2008:193). A common solution to this conundrum is the application of easily measured criteria, usually the years of formal training undergone by the learner, which despite certain shortcomings, since such aspects as learner aptitude, teaching methods, L1 or L2 status of the teacher are often ignored, seems to have become the norm for learner corpora (Granger 2012:9).

The criteria defining the content of the corpora are somewhat less tangible, but boil down to the an emphasis on open-ended continuous discourse data, aiming to collect as *natural* language as possible, with the linguist's only intrusion being whatever minimal disruption is required to collect such data, and with all data occurring in context, instead of

isolated words, phrases or clauses. According to Sinclair, learner corpora should ideally be strictly spoken, containing nothing but learners' conversations with native interlocutors as they "go about their normal business" (Sinclair 1996), but this is seldom the case: of the 122 learner corpora listed by the Université Catholique du Louvain[1] only the European Science Foundation Second Language Database (ESF) claims to consist entirely of such natural language use, while a handful of other corpora contain spontaneous language use in classroom contexts. The vast majority depend on written semi-natural language, usually in the form of open-ended elicited compositions where a specific task is given, but the learner is free to choose his or her own wording. Although the study of more specifically elicited data, such as tasks in which certain structures or words are directly requested or the learner is asked to judge whether a construction is grammatically correct, can yield valuable insights, such data is not eligible for inclusion in a learner corpus (Granger 2013:1). The overrepresentation of written, elicited data is not merely a case of researchers "not being bothered" to go through the demanding process of collecting spontaneous spoken data, but often an intentional means of keeping corpora comparable. The International Corpus of Learner English (ICLE), for example, remains one of the most-used learner corpora of English (ibid.) despite being entirely elicited, because comparability is retained by giving similar tasks to all contributors, collecting comprehensive learner profiles and keeping the number of words/learners even throughout, making it highly a highly useful tool with which to identify interlanguage features across subcorpora[2].

That the ICLE corpus lends itself so well to comparative studies is no coincidence; in fact, this comparison of interlanguages is an important part of the Contrastive Interlanguage Analysis methodology, a companion framework developed parallel to the corpus itself. While the much-employed technique of *contrastive analysis* investigates the ways in which two separate languages differ or resemble each other, CIA "does not establish comparisons between two languages but between native and learner varieties of one and the same language" (Granger 1996:43). Despite its influence, the methodology of CIA is rather incomplicated, as shown by figure 1 below: Interlanguage "E2X" is compared to interlanguage "E2Y", as mentioned above, or interlanguage "E2" is compared to native language "E1".

[1] Retrieved from http://www.uclouvain.be/en-cecl-lcworld.html
[2] Corpus collection guidelines for ICLE subcorpora provided at http://www.uclouvain.be/en-317607.html

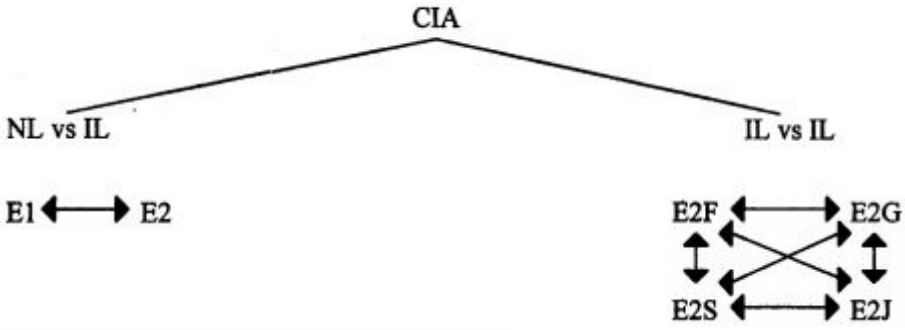


Figure 1

These E1-to-E2 comparisons necessitate a meticulous process of selection in order to ensure that the native material against which the learners are measured is suited to the task. Although comparing learners' written tasks to native conversational English could very well reveal interesting things about these genres themselves, such a comparison has less to offer in terms of interlanguage insights.

A common way of addressing the issue of comparability is to compile a corpus for the study at hand, commonly by picking elements from existing corpora (Biber & Barbieri 2007, DeCock 2004). To ensure maximum comparability across variables, entirely new corpora can be constructed to mimic the learner corpus as closely as possible, as was the case with the development of the LOCNESS corpus, the native-speaker companion to the ICLE. This corpus is closely matched for text type, writer age and experience, and although it suffers somewhat from more hetereogenous essay topics and sparser learner profiles, it remains the most commonly used control corpus for comparative ICLE studies (Hasselgård & Johansson 2011:38), enabling for example Paquot (2008) to illustrate L1 transfer at work by investigating how French learners use exemplification structures.

Some SLA specialists argue that comparing native speakers to learners regardless of how well variables line up constitutes a comparative fallacy, since learner language is, in their view, a language variety in its own right. Despite this, the use of native speaker corpora as control samples remains widespread among learner corpus researchers, who argue that although not perfect, this comparative approach is still preferable to the more intuition-based approaches of many SLA studies (Granger 2013:4). Even among those who reject these criticisms, however, there are those, perhaps most notably Geoffrey Leech (in Ädel 2006:206-7)., that question the native speaker ideal on different grounds: Is native a speaker a model

worth imitating simply by being native, or are professional writers a preferable yardstick against which to measure learners? The answer, Annelie Ädel suggests, is both yes and no. On the one hand, comparisons with professional writers are useful, as "it can be argued that professional writing represents the norm that advanced foreign learner writers try to reach and their teachers try to promote" (2006:7). On the other hand, however, foreign language essays represent a highly distinct text type, and thus comparing these texts to the output of professional writers is, as Günter Lorenz puts it; "both unfair and descriptively inadequate" (Ädel 2006:203). The ideal, then, would be to compare both, as both approaches have their merit, shedding light on different aspects of learner language, contributing in slightly different ways to our understanding of learner language. This present paper meets this ideal halfway, by using a non-professional L1 corpus as grounds for comparison, complemented by earlier research comparing the L1 corpus to an L1 corpus of published articles, in hopes of gaining not only an insight into how Norwegian users of academic English measure against their native peers, but also into the challenges common to those trying to learn the conventions of academic language.

2.3 English for Academic Purposes

The famed "man in the street" seems to have a distinct sense of what "academic" language entails. A quick, admittedly ludicrously unscientific university break-room survey revealed a shared impression among non-linguists of academic language as "heavy", "wordy", "highly specific" and far removed from everyday spoken language. In more technical terms, most scholars seem to agree that academic English is characterised by two central features that support this notion of it representing a counterpoint to spoken language. First, it is highly grammatically complex, frequently employing elaboration techniques, especially dependent clauses, a feature closely associated with the written medium, and strongly contrasting with the typical "simple and short clauses" (Hughes 1996:33-34) of spoken language. Secondly, EAP is thought to start at a disadvantage, not having the luxury of the shared situational context of speech, and must compensate with a "high degree of specificity" (Wright 2008:292), providing all assumptions and interrelations overtly in the text in order to ensure that all meanings and references are unambigously communicated. These claims are frequently repeated by scholars of EAP (Hyland 2007:284, Keen 2004:95, Li & Ge 2009:98),

but some recent studies have begun to challenge these notions: Biber and Gray (2010) conducts a diachronic study of a corpus of their own composition, and presents convincing evidence that while EAP writing is indeed elaborate and explicit, the manner in which it is so is not as traditionally assumed. Rather than relying heavily on clausal subordination, the traditional measure of elaboration, EAP instead uses embedded phrases to achieve the same effect. However, these phrasal modifiers cause a loss of explicitness: In identifying referents, a high degree of specificity is indeed retained, but the expression of ”logical relations among elements in the text” (Biber & Gray 2010:18) suffers. This style favours the professional reader, as the compact expression of meaning through embedded phrases enables an expert reader to quickly scan through texts in search of relevant passages, while having enough knowledge of the subject matter to clear up any ambiguities caused by the inexplicit style. For anyone lacking in specialist knowledge of the subject matter, these texts can easily become impenetrable, however, as they are more likely to make faulty or time-consuming inferences when encountering more opaque logical reference (Biber & Gray 2010:19).

Apart from this seeming disagreement as to what characterizes EAP, academic language is also especially interesting because it represents a genre in which being *native* gives no guarantee of being *right*. Tribble (2011) distinguishes between *apprentice* and *expert* writers: The expert texts are the peer-reviewed, professionally edited and published articles that can be found on university syllabuses, and the students reading them are expected to attempt to recreate their generic features, making their efforts *apprentice* efforts, regardless of their achieving this objective or not. Chen and Baker (2010), by comparing the apprentice texts of the BAWE corpus - to which we will return below - to the expert texts found in the FLOB corpus, confirmed that in several areas, the discrepancies between apprentice and expert texts are in fact as significant as those between native and non-native writers. Academic English, it would seem, is not only a challenge of pure linguistic competence, but also mastery of a new genre, convention and idiomaticity.

2.4 Phraseology

"You shall judge a word by the company it keeps" (J.R Firth)

While many fields of inquiry were indeed fundamentally changed by the insights a corpus can give, some were born entirely between the lines of the ever-increasing number of corpora. Among the most influential of these disciplines is the study of *phraseology*. Phraseology, as the name implies, is the study not of how a single word behaves, but of how words combine to form *phrases*, and of how new meanings and functions emerge from these combinations. Such an approach is highly conducive to the study of the aforementioned *idiomaticity* of EAP, but it is imperative to acknowledge that phraseology and the idiomaticity it investigates is no mere study of idioms in their lay sense: Discovering the mechanisms that enable us to understand that someone asking "Does the bear shit in the woods?" is not necessarily interested in the peculiarities of ursine excretion habits, but simply responding to question he perceives as entirely superfluous, is as much a study of etymology and popular culture as it is of phraseology, but the manner in which context, co-text, structure and lexis work together to create a new meaning still serves to demonstrate the synergy of co-occurrence that is the impetus for the field of phraseology. Such *idiomaticity* in its broader sense is a chief concern of many phraseologists, and the crucial feature around which early typologies were created. A.P Cowie, for example, presented his phraseological continuum in 1981, in which multi-word combinations were graded as free combinations ("blow a trumpet"), restricted collocations ("blow a fuse"), figurative idioms ("blow your own trumpet") and pure idioms ("blow the gaff"). (Granger & Paquot 2008:36) These classes demonstrate how native speakers can intuitively identify combinations that are "off", and it is this idea that Andrew Pawley and Frances Hodgett Syders's expanded on for their seminal 1983 article "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency": How is it that certain constructions are perceived as more "right" than others by native speakers, regardless of being equally grammatically correct? While their article offers no conclusive answers, its topic resonates throughout phraseology, and remains relevant even today.

In what seems to be becoming somewhat of a recurrent pattern of its own, John Sinclair's contributions to the field in the 1990s turned phraseology entirely upside down. In line with the ideas of corpus drivenness discussed above, Sinclair suggested an approach to phraseology that was based primarily on objective criteria, with frequency once again at the helm. Sinclar emphasised the idea of *collocation*, the tendency of words to co-occur, as a key

component in how language works, quite contrary to popular opinion at the time: the traditional idea, on which virtually all grammars are based, he claims (1991:110), is that a speaker, once a unit - be it a word, phrase or clause – is uttered, picks his or her next unit from the entire breadth of the lexicon, only restrained by the grammaticalness of the unit – The *open choice principle*. The alternative proposed by Sinclair is the *idiom principle*:

"The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. […] However it arises, it has been relegated to an inferior position in most current linguistics, because it does not fit the open-choice model" (Sinclair 1991:110)

Although not the first to point out the importance of recurrent language, with Dell Hymes noting as early as 1968 that "a vast portion of verbal behaviour…consists of recurrent patterns, of linguistic routines" (Conrad & Biber 2004:56), Sinclair was hugely influential in certain circles when it came to elevating the status of such idiomaticity to "at least as important as grammar in the explanation of how meaning arises in text"(Sinclair 1991:112).

Despite these highly influential contributions, however, phraseology today is no monolithic structure: it has no definite starting point, no seminal article from which all later studies have grown, no universally held truths. This freedom to approach a study from any angle imaginable is certainly a forte of the field, but it can also be cause for some headaches for any fledgling phraseologist, as the flip side to this diversity is a complete lack of agreed-upon terminology. Phraseology is, as A.P. Cowie laments, "bedevilled by the proliferation of terms and by the conflicting uses of the same term" (1998:210). To further complicate matters, the differences are often minute: The immediate distinctions between a *recurrent word combination* (Altenberg 1998), *phraseological unit* (Fiedler 2007), *lexical bundle* (Biber et al. 1999) cluster (Scott and Tribble 2006) and *n-gram* (Stubbs 2007), are fuzzy, but to conflate them all could still lead to erroneous conclusions. Therefore, it is imperative to the following discussion that we pledge our allegiance to a set of terms, ironic as it may be after having spent quite some time extolling the virtues of untarnished corpus-drivenness. The terms and methodology used may well be borrowed, and thus presents a pre-conceived set of ideas, but it lies outside the scope of the present paper to entirely reinvent the field of phraseology, and the methodology presented below is soundly enough rooted in objective criteria to retain a corpus-driven character.

The methodologies applied to tackling mutli-word sequences vary greatly, but six variables are commonly applied in some manner: Fixedness; idiomaticity; frequency; length of sequence; completeness, be it syntactic, semantic or pragmatic; and intuitive recognition by language native speakers (Conrad & Biber 2004:57). The order in which these features are given priority will affect the outcome of the study: A study of collocation, for example, aims to identify semantic networks between words, and therefore emphasises frequency, semantic completeness and two-word relationships, while fixedness, idiomaticity and native speaker recognition is disregarded, since the meaning of a collocate can be deduced from its parts regardless of intervening words, and native speaker recognition is overridden by the statistical evidence. A study of idioms, by contrast, puts completeness, idiomaticity, fixedness and native speaker recognition first, but is relatively uninterested in frequency and sequence length, since well-known idioms are not necessarily very frequently used and come in all shapes and sizes. In the present paper, the term *lexical bundles* will be used, along with the framework in which it has been applied. The term first appeared in the *Longman Grammar of Spoken and Written English* (Biber et al 1999:990-993), which defines lexical bundles as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse". By this definition, then, primacy is given to frequency and fixedness, while the emphasis on discourse implicitly requires a sequence length of three or more words, "since many [two-word sequences] are word associations that do not have a distinct discourse-level function" (Conrad & Biber 2004:58).

Biber, who has continued to work within the framework, usually in the form of collaborative articles (see Conrad & Biber 2004, Biber & Barberi, 2007, Biber, Conrad & Cortes 2004), further refined the methodology by identifying the various functions the bundles serve in discourse: Bundles that refer directly to physical or abstract entities or to the textual context, are labelled *referential expressions*. Those that "reflect relationships between prior and coming discourse" (Biber, Conrad & Cortes 2004:384) are labelled *discourse organizers*, while *stance expressions* are bundles that express attitudes, assessment or modality in general. All three categories contain several subcategories, embellished upon in the "Methods" section below, but it is important to note that these categories are, as pointed out by Ädel & Erman (2012:89) not without their problems, as the criteria for belonging to a (sub-)category are relatively loosely defined, and bundles can sometimes belong to several categories. Despite these imperfections, however, Biber's taxonomy will be applied to the

present study, with explanations offered if difficulties should arise, in order to facilitate comparisons with earlier studies performed on native and non-native language (Chen & Baker 2010, Ädel & Erman 2012) as well as academic English (Biber & Barbieri 2007).

Biber's collaborations, as well as "independent" efforts (Chen & Baker 2010, Ädel & Erman 2012) have investigated a fairly diverse set of corpora and texts, ranging from academic to conversational English, all support Sinclair's call to elevate the status of idiomatic language. While unsurprisingly heavily outnumbered by non-recurrent sequences, lexical bundles are still pervasive, accounting for 20% of the LGSWE material (Biber et al 1999:994), and as much as 28% of conversational material (Conrad & Biber 2004:62) even with fairly conservative cutoff points. One of the strengths of the approach lies in its inclusion of structurally incomplete units. Conrad and Biber suggest that many of these have been overlooked in previous research because an incomplete unit is not as easily recognized by pure intuition, and furthermore make a strong case to suggest that these units, despite being structurally incomplete, are nevertheless important building blocks of discourse commonly used by speakers and writers (2004:69).

When investigating learner language, phraseology is in its element, as it allows for investigation of patterns of contextual mis-, under- and overuse, which several studies have shown to be as salient a feature of such language as purely morphological errors (Granger 2013:4). Norwegian learners, for example, despite being familiar with *indeed*, have been shown to be somewhat more reluctant to use it, with *indeed* 62% being more frequent in LOCNESS than the ICLE-NO (Hasselgård & Johansson 2011:39). Such an approach to learner corpora allow us to identify situations where learners resort to *avoidance strategies* because they are not entirely comfortable with the ways in which a word or construction is used (Granger 2012:10), and even better, allow us a unique opportunity to investigate *what is not there*.

3 Material

3.1 VESPA:

The "Varieties of English for Specific Purposes dAtabase" (VESPA) began development in January 2008 at the Université Catholique the Louvain under the direction of Dr. Magali Paquot, aiming to collect examples of English used for highly specific – in this case academic – purposes, by L2 users from as wide a background as possible. Leading by example, the UCL team began development on a corpus of academic texts written by French-speaking students, and contributions soon came from other universities, with German, Spanish, Turkish, Italian and Norwegian L1 subcorpora currently in various stages of completion.

Collection guidelines for these subcorpora are provided, both for what data to collect, and in what format they are to be collected. The latter of these are quite rigid, with emphasis placed on the collection of comprehensive learner profiles and text formatting. The former, however are quite general in order to facilitate continued contributions to the project, but narrow enough to ensure that the essence of the corpus remains uncompromised. Firstly, all work must be the student's own. Reference tools are allowed, but no third party assistance, including revisions based on teacher feedback are acceptable. Secondly, all data must be representative of its discipline and characteristically *academic*, discouraging contributors from giving argumentative, descriptive or narrative tasks.

Thirdly, all texts must be at least 500 words, but very long texts are acceptable, such as Master's theses. At this point a certain disharmony between criteria 1 and 3 becomes evident, as such theses are frequently the result of an extensive editing process in which a supervisor is involved, thus invalidating the second criterion. For the purposes of this paper, however, this slight inconsistency proves wholly unproblematic, as no Masters' theses are included in the Norwegian subcorpus compiled by Prof. Hilde Hasselgård and Dr. Signe Oksefjell Ebeling at the University of Oslo[3].

The VESPA-NO corpus currently consists of 422 texts for a total of 555,733 words, distributed across three disciplines; business, lingusistics and literature. The relatively limited sampling of the corpus necessarily limits the scope of the study, as a study of merely three

[3] VESPA corpus guidelines are available at http://www.uclouvain.be/en-cecl-vespa.html

disciplines could hardly claim to uncover any truths of academic writing in general. Therefore, I have opted to omit the "Business" section because it is likely to represent a rather different set of writing conventions than its humanities counterparts, potentially making the final results a somewhat haphazard collection of bundles from each field. A conflation of the linguistics and literature disciplines, however, is well suited to the task at hand, as these two disciplines are taught as a single subject in Norwegian schools, often with little explicit distinction made between the two. This continues at the university level, where students of English are expected to participate in both linguistic and literary courses. Including both disciplines, then, will provide us with a sample that lies closer to how Norwegian students are in fact using English in an academic context than simply choosing one or the other. It is, however, important to note that the literature section of the corpus is rather small compared to the linguistics section, as table 1 shows. As a consequence, any specific generic idiosyncracies that surface are likely to belong to the linguistics section, but the literary texts will still contribute to a more appropriate representation of the totality of how Norwegian students of language and literature use EAP.

Table 1. VESPA sample

	Texts	Words
Linguistics	320	475018
Literature	12	18285
Total	332	493303

3.2 BAWE

The British Academic Written English (BAWE) corpus was developed as a part of the "An Investigation of Genres of Assessed Writing" project, a collaborative effort between the Universities of Warwick, Reading and Oxford Brookes with the joint goals of identifying "the characteristics of proficient student writing produced for degree programmes in British universities» (Nesi et al. 2008:2) and establishing both genres and sub-genres in which to properly categorise these various texts. The aim of the corpus was to facilitate this investigation, and thus it consists of a wide variety of texts produced by both graduate and undergraduate students from British universities across most disciplines, ranging from leisure management to cybernetics. While similar discipline-specific samplings had been made for individual studies, the BAWE corpus was the first publically available large-scale formal corpus compiled from such material (Alsop & Nesi 2008:72).

Similarly to the VESPA, BAWE consists of mandatory assignments submitted by students in partial fulfilment of module requirements. Regardless of whether or not the assignment would count toward the final grade given, all submitted texts are held to a certain standard of proficiency, rejecting all texts failing to meet the "departmental standard" (Alsop & Nesi 2008:71). Although the intentions of its creators were for the corpus to contain an equal amount of submissions from each level of study, the collection process proved an arduous task, with highly varying amounts of data collected for each discipline and level, resulting in a corpus somewhat skewed toward lower-level texts in terms of distribution. With the higher-level texts somewhat longer on average, the word count remains fairly balanced when investigating the corpus as a whole, but certain disciplines are heavily underrepresented. (Alsop & Nesi 2008:79-80).

The present paper will be using only the Linguistics section of the corpus. Adding the "English" section was initially considered, but the idea was rejected, as a pilot study indicated that this would skew the sample heavily toward literature, making it ill-suited for comparison with the linguistics-dominated VESPA sample. Opting to investigate only the Linguistics subcorpus also reduces the impact of the aforementioned internal balancing issues: perhaps unsurprisingly, seeing as the amount of submissions to the corpus depended entirely on student interest, the project garnered quite some interest among linguistics students, making the linguistics section a reasonably balanced affair. In light of this, the metadata supplied for student level and assignment grade will be disregarded and the corpus will be investigated as a whole, with an important exception: Only students with English L1 will be included, since the purpose of using BAWE as a control corpus would be severely undermined by the inclusion of non-native data. With all non-native texts removed, our sample is pared down somewhat, leaving us with a total of 181, 813 words distributed across 75 texts.

3.3 Comparability

Comparable corpora, as defined by Johansson, "consist of original texts in each language, matched by such criteria as time of composition, domain, genre, intended audience etc" (1999:5), and although Johansson in this definition is referring to sets of corpora in different languages, the definition is equally useful applied to varieties of a single language. According to these criteria, the VESPA/BAWE pairing seems well-matched, with all variables lining up fairly neatly. All contributions are from roughly the same period, the demographic under

investigation is similar in both samples, and although there are slight generic variations, all contributions are written with the same purpose, that of demonstrating one's capability in an academic field. The data collection regime is similar, with only written data collected. The only major deviation between the two is the lack of a "departmental standard" for VESPA contributors, but with BAWE's primary function in the study being that of control sample, this is hardly detrimental to the results, since the basis of such a comparison is the assumption that a proficient native is a suitable role model for learners. The two corpora also differ somewhat in terms of size, with the BAWE sample only half the size of the VESPA sample. For the fairly narrow,practically single-discipline scope of the present study, however, the size of both corpora is adequate, with a broad enough range of contributors to avoid severe overrepresentation of individual writer idiosyncracies, and enough text in total to "dilute even the longest texts" (Sinclair 2005:7) Crucially, both corpora are annotated along the same lines and using the same TEI-compliant tagset, developed for the BAWE project by Signe Ebeling and Alois Heuboeck (2007), allowing for similar variables to be applied when conducting searches within the two corpora. With these tagsets applied, the effective size of the corpora is reduced, with the final count for VESPA being 380,109 words, while BAWE is comparatively unscathed at 165,239 words.

In sum, BAWE seems to provide a suitable native control sample against which to measure the VESPA corpus.

4 Method

4.1 Framework

The present study is largely inspired by Biber's work with lexical bundles, and will be following the method initially outlined in Biber et al. (2003), taking into account some recent adjustments to and criticisms of the approach, both from Biber's further work in the field (Biber et al. 2004, Biber & Barbieri 2007, Biber & Gray 2013) and others (Ädel & Erman 2012; Chen & Baker 2010).

Where the orthodox method of sample selection for lexical bundle studies is the establishing of a cutoff frequency, this approach proved less suitable for the present purposes, as far more results were returned from VESPA than from BAWE. This is interesting in itself, and this discrepancy in frequency will indeed be the subject of further discussion below, but for comparative purposes, a fixed number of sequences from each side is better suited to our needs, and the present study thus opts to extract the 250 most common sequences of three or more words from each corpus. Although the present study flouts Biber's criterion of a cutoff frequency, the criterion of *distribution* is retained: all sequences must be present in 5 or more texts, in order to "guard against idiosyncratic uses by individual speakers or authors" (Biber et al. 2004:75).

Four pieces of software are used in the extraction, ordering and visualisation of lexical bundles and their distribution: WordSmith Tools (Scott 2001), Microsoft Excel, AntConc and Filemaker Pro. The corpora, both provided through the University of Oslo in XML format, are loaded into WordSmith, and a list of relevant tags is added to the exclusion list in order to ensure that no direct quotes, references or other material not originally written by the contributor are included in the returned results. Using the WordList, Index and Cluster Calculation tools, recurrent word combinations are identified, extracting all clusters of 3+ words. The frequency threshold is set at 10, simply in order to be high enough to facilitate quick processing and low enough to return at least 250 clusters, which are then then exported to Excel, where relevant additional data is calculated, including a normalized frequency per 100,000 words and distribution level across texts. All returned clusters met distribution criteria, and thus no clusters were removed, leaving us with 250 *lexical bundles*. These are

then formatted for export to FileMaker, where an interface is set up to categorize all clusters after their function in discourse.

4.2 Classification scheme

The three main categories of stance expression, discourse organisers and referential expressions outlined above is embellished upon by Biber and Conrad (2004:65-66), adding subcategories for each function, a distinction between academic prose bundles and conversational bundles, as well as adding a separate category for conversational functions. No such conversational bundles are found in VESPA, suggesting already at this stage that the VESPA contributors have a solid grasp of the divergent conventions of written and spoken language. The subcategories and criteria for the three categories found in VESPA are as follows:

4.2.1 Discourse organizers

Discourse organizers "identify a logical relationship between a prior discourse segment and the subsequent discourse" (Biber & Conrad 2004:81). They act as guides for the reader, as preparatory aids, indicating how the following information is to be interpreted. They can introduce a new topic or series of arguments (1), signal that the coming information is a paraphrase of previously given information or a narrowing of focus(2) or signal a shift towards another aspect of the topic under discussion (3).

Topic introduction

(1a)"**First of all**, it is important to link the title to the text itself." VESPA UIO0112-LIN-02

(1b)"**In the case of** verbs, a large majority of these are characteristically found at the head of verb phrases.." BAWE 6120c

Topic clarification:

(2a)**"In other words,** "there" introduces a subject, as happens to be with our example: the unknown "man in a sports car" is introduced for the first time." VESPA UIO0058-LIN-01

(3a)**"In terms of** social skills**,** and her non-academic studies, Sunny seems to have acclimatised well to her new life in Britain." BAWE 3118b

Topic elaboration

(3a3)“Halliday, **on the other hand**, only includes the existential 'there'.”VESPA UIO0019-LIN-02

(3b)“**As a result of** this global spread of English, Cheshire's Anglo-centric description of English is no longer relevant.” BAWE 6020d

4.2.2 Stance expressions

Although it can be argued that all texts are inevitably implicit expressions of the writer’s attitudes and positions, the lexical bundle framework eschews such philosophical questions for a more pragmatic approach, labelling as “stance expressions” only those constructions that make these views explicit to the reader. This is realized largely by four types of bundles:

Epistemic:

(4a) “This reflects **the fact that** the part of the text that is below line 13 contributes to establishing the closer contact between the writer and the reader which helps to manipulate the final choice of the reader.” VESPA UIO0104-LIN-01

(4b)“In considering the 'sex-exclusive' idea, **it is clear** by observing language use in British society today that men and women are in fact using the same language.” BAWE 6126d

Obligation/directive:

(5a) “In the end **it is important to** remember that you can never fully recreate a literary piece into film, because of the details and length, and since it is such a subjective experience for every individual to read a book.”VESPA UIO0185-LIT-01

(5b) “However**, it is important** to consider that class is not fixed, for example, a woman may have been born into a low social class, but marry a man from a high social class and would therefore have to take his class position.” BAWE6042a

Ability:

(6a) “In those cases, **it is possible** to understand the main content by reading only the Themes.” VESPA UIO0086-LIN-03

(6b)“Also**, it is possible** to involve students in negotiating topics or outcomes, a major feature in the TABASCO project” BAWE 3127a

Prediction/inference:

(7a)“These errors are **likely to be** interlingual errors.” VESPA UIO0149-LIN-03

(7b)“Given the ending of the song **this is unlikely** to be a coincidence.” BAWE 6018a

4.2.3 Referential bundles

The final category, the “workhorses” of an academic text, comprises the variety of constructions that serve to structure, embellish and situate an argument. They do so by referring to objects both physical and abstract, as well as internally in the text, and are as such labelled *referential bundles*. Referential bundles are ubiquitous in academic writing, represented by a wide variety of grammatical constructions. Instead of identifying them by the attributes they possess, a process of elimination is often equally effective: a bundle that does not directly contribute to the organisation of a text or explicitly reflects the attitudes of the writer is likely to serve a referential function. The most common such functions are given below:

Attribute specification:

(8a) “All in all **it is a** fairy tale ending” VESPA UIO0192-LIT-01

(8b) “**It is a** noticeable trend within Fig. 1 that all the FS pronouns have appeared less frequently than their FP counterparts.” BAWE 6048b

Text deixis:

(9a) “These four all occur **in the first** four lines, and it is safe to say that they do not exactly make for a good first impression.” VESPA HIOF0005-LIN-02

(9b) “There are three pauses **in the first** half of this line.” BAWE 6009b

Personal deixis:

(10a) “Text 1 wants to bring the facts **to the reader** while text 2 is a story which most likely isn't true and just for entertaining.” VESPA UIO0160-LIN-01

(10b) “It is true **they do not** have the same linguistic difficulties, but may lack some of the academic skills already practised by and familiar to many EAP students.” BAWE 3118b

Time markers:

(11a) “In the example above, it seems that the process of treading water is happening **at the same time** as Bernard says he is sorry.” VESPA UIO0001-LIN-05

(11b) “The examination of extended discourse has provided yet more evidence that language development continues after five years, whilst **at the same** time also showing the difference in development before and after five.” BAWE6020b

Framing attribute:

(12a) “This brings with it a radical change in the story **as a whole**, and can perhaps be a bit disappointing to lovers of Defoe's original work.” UIO0185-LIT-01

(12b) “There is no doubt that the study of quantifiers, particularly **in the context** of negative clauses has presented logicians and linguists alike with a number of perplexing and contradictory phenomena.” BAWE 6038d

4.2.4 Content-specific bundles

For the present study, a fourth category is added, inspired by Chen and Baker (2010): “Content-specific bundles”. This is a necessary adjustment to the method due to the size and single-discipline focus of the corpus at hand allowing highly specific clusters to dominate in terms of frequency. This is the case with entire clusters such as “the present progressive” or “Australian and New Zealand English” as well as single, repeated instances of clusterings around words. An example of the latter is “text”, which is extremely frequent, likely because assignments in both linguistics and literature commonly revolve around text analysis in some form. Scholars such as Stubbs and Barth (2003) have demonstrated that studies of content-specific bundles can yield valuable insights in the fields of lexicography, linguistic forensics and stylistics,while Biber himself (with Gray, 2013) has turned to examining the *frames* in which such frequent words occur, but such endeavours are sadly outside the scope of the present study, and these content-specific clusters will be omitted from the further investigation.

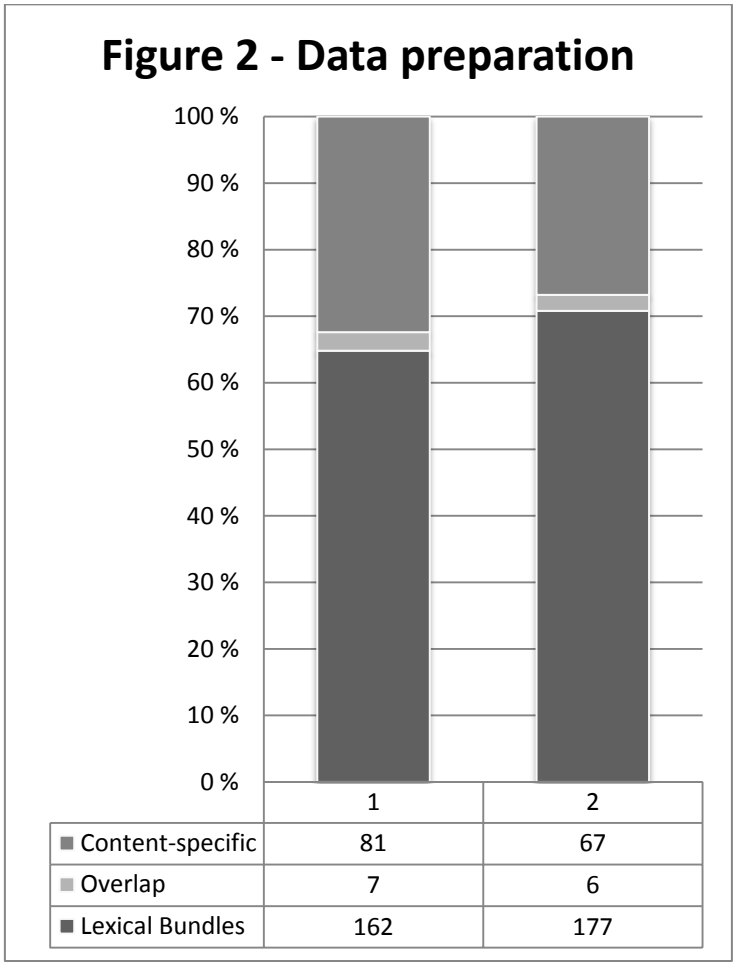
4.3 Data preparation

The role played in discourse by each bundle is determined by investigating each bundle in context by loading the relevant corpus into AntConc, a concordancer similar to WordSmith, but with an interface more suited to repeated queries. For multi-functional bundles the most common usage was given precedent. It must be noted that the categories were initially presented as a "preliminary" categorisation (Biber et al. 2004:83), and despite the subsequent studies carried out since their introduction they have yet to evolve into a clearly delineated system, so a certain amount of intuition is inevitably at play, leading to some potential inconsistencies, as remarked by Ädel and Erman (2012:89). This is a weakness of the framework that the present study accepts, but the damage done is marginal as long as the same criteria are applied and the same judgments made for both corpora, since it is the comparison between these two that is the chief goal of the study. There are also instances of bundles that can function in more than one manner, as with "can be seen", which acts as a referential expression in (13), but a stance marker in (14). Again, frequency is given primacy, with each bundle categorised according to its predominant function, in line with what seems to have become common practice for such studies (Ädel & Erman 2012, Chen & Baker 2010, Biber et al. 2004)

> (13) "Examples of this **can be seen** in lines 16, 38 and 39" BAWE6062a
>
> (14) "This **can be seen** as an interlingual error in the way that she may have thought of the Norwegian word and then presumed that the English then takes the definite article, when it does not.." VESPA UIO0180-LIN-02

The final step of the data preparation, borrowed from Chen and Baker (2010:33), is the removal of overlapping word sequences in order to avoid misrepresenting the importance of certain clusters: "The other hand" and "on the other", for example, are both frequent sequences in both corpora, but a concordance search reveals that nearly all these occurrences are as part of the sequence "on the other hand". These are removed from the results, as they are essentially noise generated by the computerised extraction, while "on the other hand" is kept in. Hypothetically, there could have been instances of such overlapping patterns

occurring often enough in separate contexts to warrant inclusion among the most frequent bundles, but this was not the case in the present study. The bundles are culled from the final listing after being exported to Excel. The results of the data preparation process are seen in Figure 2.



All further ordering, calculation and visualisation is done using the conditional formatting, sorting and graph functions of Excel.

5 Data and analysis

5.1 General characteristics

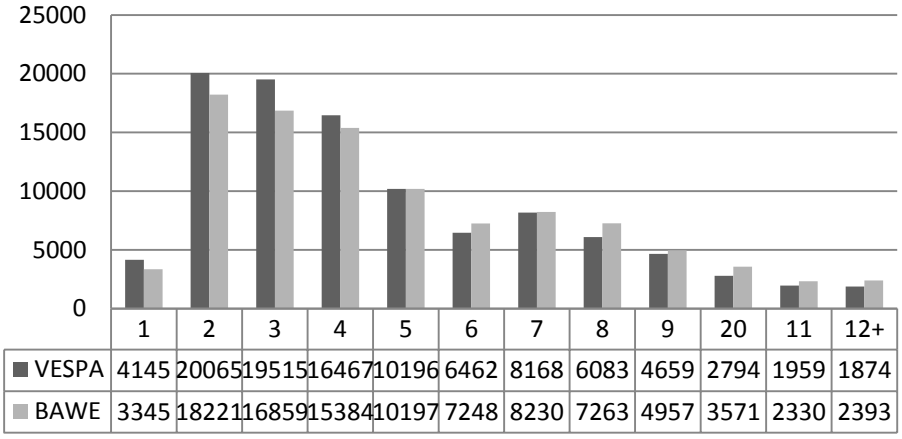
Although they are at best tangentially related to the lexical bundle framework within which the majority of the present study will be carried out, three simple statistical measures have been included in the study. Despite being decidedly un-phraseological in their approach, these can still contribute to the description of certain salient features of native and non-native EAP use, by briefly investigating the building blocks of discourse: words and sentences.

At a fundamental level comparing average word and sentence lengths can provide some basic indicators of how the two corpora approach academic language. L2 users tend to use slightly shorter sentences, with an average length of 20.76 words per sentence, 17% shorter than their native counterparts, who average at 24.88.

Word length preferences lean toward shorter words in VESPA, with one-, two-, three- and four-letter words approximately 10% more common than in BAWE, and 6+ letter words equally underrepresented, as seen in fig 3.

Figure 3 - Word length

In letters. Per 100k words



	1	2	3	4	5	6	7	8	9	20	11	12+
VESPA	4145	20065	19515	16467	10196	6462	8168	6083	4659	2794	1959	1874
BAWE	3345	18221	16859	15384	10197	7248	8230	7263	4957	3571	2330	2393

A third statistical measure, Type-Token Ratio (TTR), is calculated by dividing the the unique words used (*types)* in a text with thetotal number of running words (*tokens)*, indicating the size of the vocabulary employed by the writer – the text's vocabulary richness - with a

percentage value. A text where no word is repeated scores 100%, while a 100-word text where 30 words are repeated gets a TTR of 70%. As a text grows longer, more and more words are repeated, and the TTR value drops rapidly, usually stabilizing at around 4% for the larger, multi-genre corpora. With the highly specialized vocabulary that can be expected of a single-discipline academic corpus in mind, the 2.6% TTR of the VESPA corpus does not seem unnaturally low, but when compared against BAWE, VESPA is less favourably portrayed: Even when a Standardised Type-Token Ratio analysis is applied, dividing the text into 1000-word sections and calculating the average TTR in order to compensate for the discrepancy in size between the two corpora, the BAWE values are twice as high as in VESPA, with an STTR of 5.8% compared to VESPA's adjusted 2.5. Interestingly, though, the number of types is fairly similar, as seen in table 2, and a WordSmith wordlist comparison shows that the overlap in types is significant, as only 156 of the 9142 types in BAWE are not found in VESPA, thus giving a lexical overlap of 98%. Classifying these words according to their word class (Table 2) reveals that almost a third are proper nouns, most commonly in reference to scholars, while the remainder are all highly specific verbs, nouns or adjectives, largely from the fields of phonetics ("obstruent", "fortis") and neurolinguistics ("aphasia", "autistic", "hemisphere"). This classification was carried out manually, since no such tags are built into the corpus.

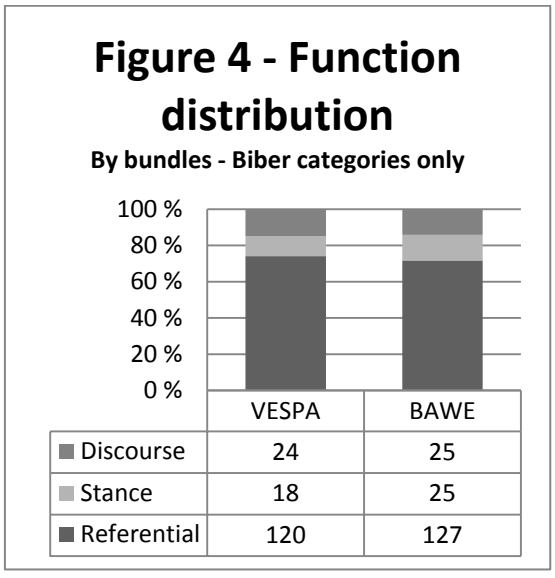
Table 2 - Words unique to BAWE	
Proper nouns	55
Adjectives	26
Nouns	70
Verbs	2
Noise	2

The presence of such content-specific words is not necessarily indicative of more than a relative overrepresentation of phonetics and neurolinguistics in the BAWE sample, and we could reasonably assume that the majority of "missing" words would have been included had these disciplines been equally present in VESPA. What is striking, however, is the complete absence of functional words unique to BAWE, suggesting that the two samples have a shared core vocabulary of general purpose words. This supports the initial assumption that it is not

the lexis itself that sets interlanguage apart from native language, but rather the ways in which lexis interacts with other aspects of language in varying degrees of native-like idiomaticity.

5.2 Lexical bundles and functional distribution

The distribution across discourse functions of the identified bundles - shown in Figure 4 - is remarkably similar for both corpora. An overwhelming majority of the bundles serve a referential function, while stance expressions and discourse organisers account for roughly 15% each. This ordering is similar to that found in previous studies of academic learner language, with the stance-to-discourse marker ratio around 1:1 in all sets, and the referential expressions dominating the sample. The extent to which the referential expressions dominate is far greater in the present study, however, where this function is more than 50% more common than in both Chen and Baker and Ädel and Erman's studies.



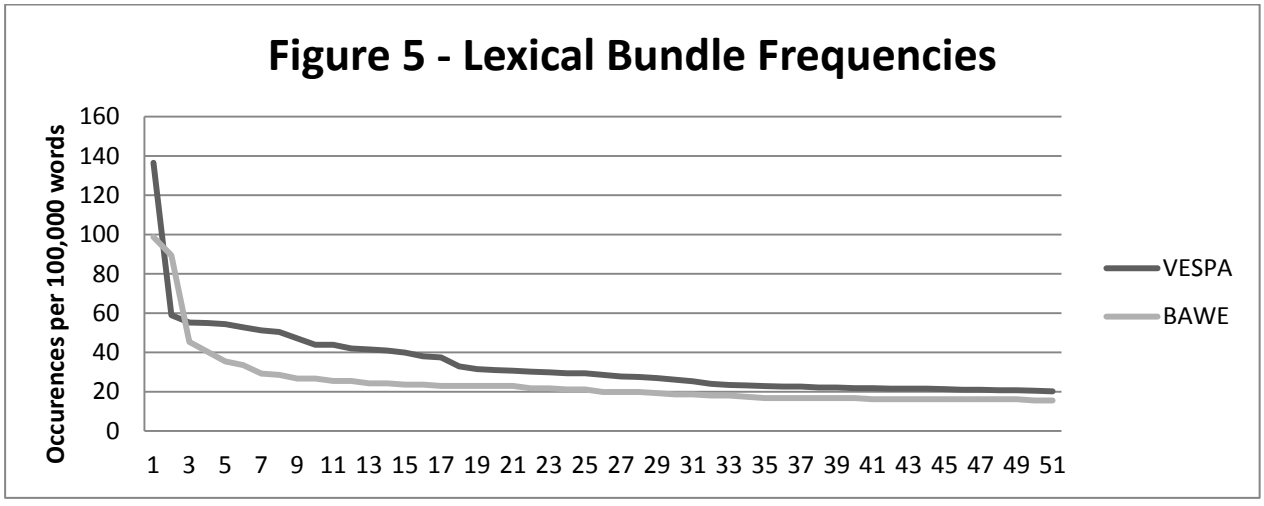
This discrepancy may be due to several factors. First, earlier studies have focused exclusively on four-word bundles, which might appear to be a more productive format for stance and discourse functions, with several of the most frequent bundles in both these categories consisting of four words: "we can see that" and "on the other hand". Secondly, since the present study is concerned with linguistics and literature studies data, the language is flavoured by the meta-discussions that pervade these fields, meaning that a perhaps larger portion of multifunctional bundles are predominantly referential instead of discoursal, as

many references to texts, words or paragraphs refer to the language under investigation rather than act as guides within the text itself. In addition, the very format of many tasks assigned in the discipline at hand are often fact-oriented, and as such do not require the nuanced hedging of other disciplines. While one would do well to suggest or indicate probable causalities in for example social sciences, a student of phonetics stating that the English letter “B” merely “seems” to be realized as a lenis bilabial plosive is not necessarily equally rewarded for his diplomacy.

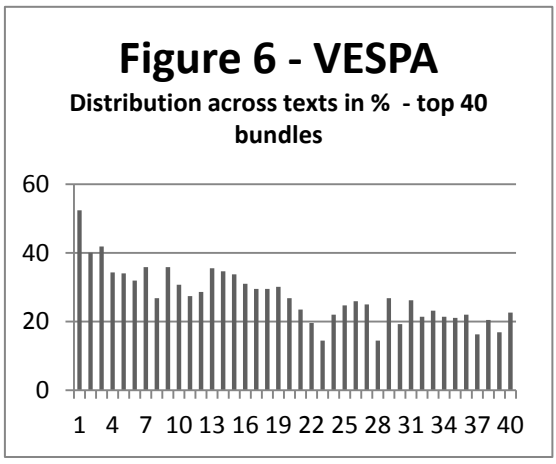
Finally, the ever-returning issue of the lexical bundle method’s rather general category assignment criteria is likely to affect the outcome somewhat. The present study has interpreted the term “discourse organiser” perhaps more strictly than others, opting to label as such only bundles that explicitly contribute to the organisation of a text, while labelling more sentence-internal reference and deixis as referential.

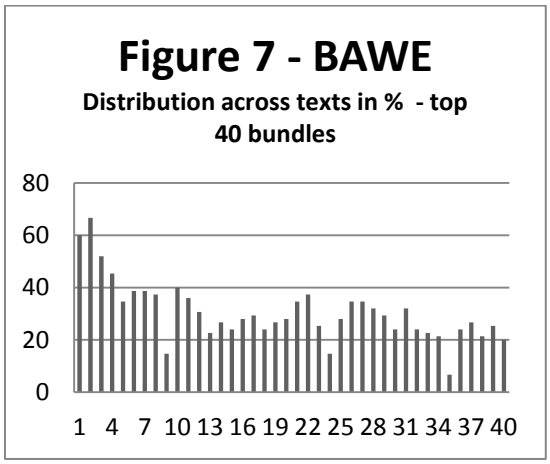
5.3 Lexical bundle frequencies

While the lexical bundles identified seem to serve largely similar functions, the frequency of their occurrence varies somewhat between the two corpora. The trend is similar on both sides, as illustrated by fig. 5, starting with a few extremely common bundles and dropping sharply before stabilizing, all within the span of 40 bundles, but the drop is far more abrupt in BAWE: only six bundles are more frequent than 30 per 100 000 words, while 22 bundles are equally frequent in VESPA. Similarly, the combined normalised frequency for the top 40 bundles is more than 50 percent higher in VESPA, account for nearly 45% of the entire sample.

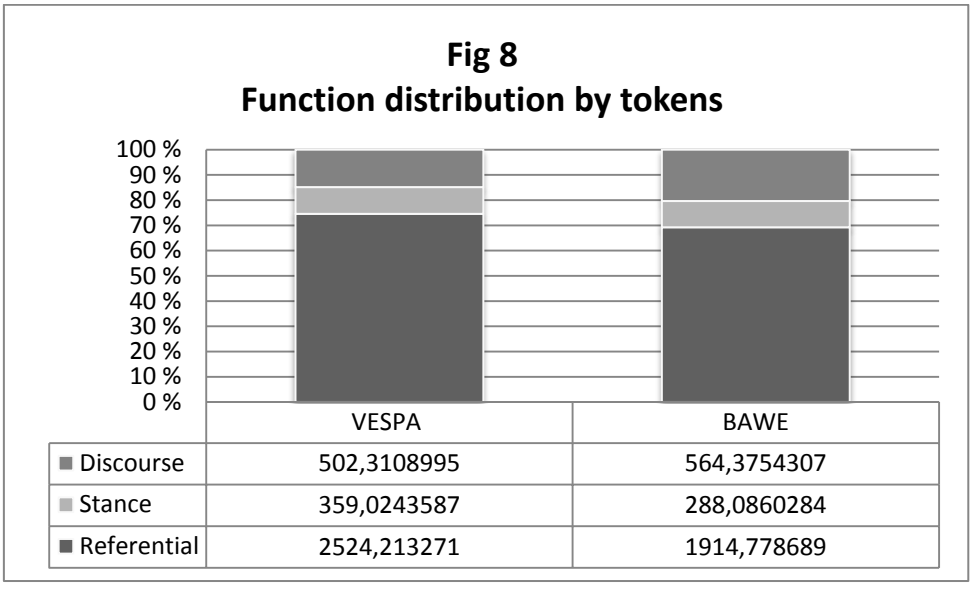


The distribution across texts for each bundle, seen in figures 6 and 7, is fairly wide. Since both corpora consist of a relatively large number of fairly short texts, this is to be expected: after all, a bundle only found in a small number of texts would have to be *extremely* frequent in order to appear among the top 250. The average distribution is again slightly higher for VESPA.





In terms of frequency, the distribution across functions for all bundles reveals an interesting pattern for VESPA contributors, as the initially under-represented stance category now matches the BAWE sample in sheer frequency. This suggests that VESPA contributors are in fact giving expression to their personal stances in their texts as often BAWE contributors, but that they do so with a more limited set of expressions at their disposal.



	VESPA	BAWE
Discourse	502,3108995	564,3754307
Stance	359,0243587	288,0860284
Referential	2524,213271	1914,778689

These factors combined indicate that the VESPA contributors have indeed appropriated a seemingly native-like propensity for using and re-using especially productive phrases, but display less variation in their use of these, with a few top-tier bundles used excessively compared to their native counterparts.

5.4 Shared bundles

Even though the two corpora contain very similar texts, the lexical bundles found in VESPA overlap only partially with those of BAWE, as shown in table 3, with a total of 78 VESPA bundles used also by native writers. In terms of frequency, these bundles account for 60% of the total, despite making up only 44% of the bundles. This overrepresentation is concentrated predominantly around the very most frequent bundles, with all the ten most frequent VESPA bundles reoccurring in BAWE, albeit not always in equal proportion. Not only are L2 users using lexical bundles in their writing, but these bundles are in fact more often than not the very same bundles as those used by native speakers, suggesting that these constructions are perceived as more native-like and "correct" by the writers employing them, since they are either used more frequently or by more writers than their non-attested counterparts. A shared bundle is not necessarily used equally often in both corpora, and as table 4 demonstrates, the discrepancies can be quite severe, with some bundles more than four times as frequent in VESPA as in BAWE.

Table 3 - Shared and unique bundles

Shared:	78
VESPA Uniques	98
BAWE Uniques	109

Table 4: Overused in VESPA

Function	Bundle	BAWE Freq.	VESPA Freq.	VESPA/BAWE ratio
Referential	IN THE FIRST	8.7	47.1	5.4
Discourse	ON THE OTHER HAND	11.2	43.9	3.9
Referential	PART OF THE	16.1	54.9	3.4
Stance	THE PURPOSE OF	9.3	29.4	3.2
Referential	FOUND IN THE	9.9	29.4	3.0
Referential	THIS IS A	16.1	42.0	2.6
Discourse	WE CAN SEE	8.7	22.6	2.6
Referential	AS IN THE	8.7	21.5	2.5
Referential	THE USE OF THE	15.5	37.4	2.4
Referential	BETWEEN THE TWO	11.2	27	2.4

A fair share of these bundles are likely owed to the nature of the assignments given: "In the first" is not unexpected from texts frequently using textual deixis to refer to a passage in a text

under investigation, nor is *the purpose of; this is a; as in the* or *the use of the*. It is evident, however, that VESPA contributrs are more prone to using the active construction "we can see" rather than the passive constructions preferred in academic writing. (Biber et al. 1999:937). The overuse of the discourse organizer *on the other hand striking* is striking, and will be the subject of a more in-depth discussion below.

VESPA *underus*e, however, is not as severe in terms of frequency, but there are a small number of bundles that are far more commonly used by the native writers, shown in table 5.

In stark contrast to the more subject-specific bundles in table 4, however, all but one of these bundles –*that there is* – are typically "academic" phrases. In fact, Liu (2007), by investigating and combining the results from a variety of academic corpora, was able to identify the phrases most frequent in academic writing, a list that confirms that these four bundles are in fact crucial elements of an academic style, all occuring more than 100 times per million words across Liu's sample.

Table 5: Underused in VESPA:

Function	Bundle	BAWE Freq.	VESPA Freq.	BAWE/VESPA Ratio
Referential	A NUMBER OF	23.6	12.7	1.9
Referential	BE ABLE TO	29.2	16.7	1.7
Referential	DUE TO THE	35.4	21.3	1.7
Discourse	IN ORDER TO	89.4	55.2	1.6
Referential	THAT THERE IS	23.0	15.6	1.5

The distribution of functions of these shared bundles remains fairly equal, as fig. 9 indicates, but sees a slight underrepresentation of stance bundles, contributing to the emerging motif of Norwegian L2 users struggling with the expression of stance, demonstrating an over-reliance on a relatively limited set of such constructions, a minority of which are in fact used by native speakers.

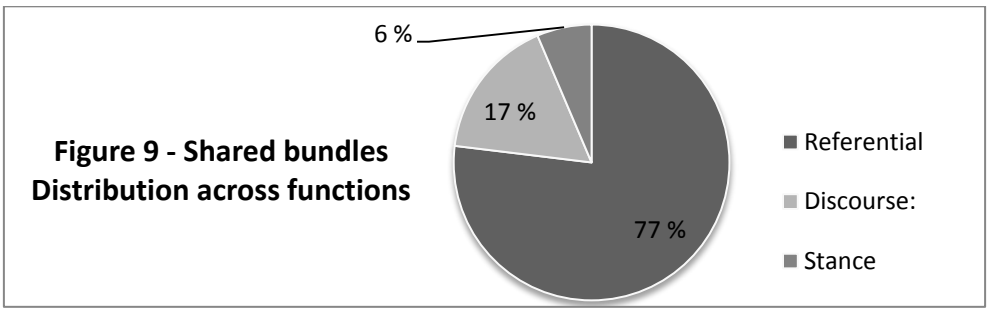


Figure 9 - Shared bundles Distribution across functions

5.5 Unique bundles

Table 6 -VESPA Uniques

Rank	Bundle	Norm. freq	Function
11	SEEMS TO BE	43.9	Stance
15	A LOT OF	39.9	Referential
16	THE MEANING OF	38.0	Referential
17	THE USE OF THE	37.4	Referential
18	LOOK AT THE	32.9	Discourse
19	IT IS A	31.5	Referential
20	SOME OF THE	31.0	Referential
21	IT IS NOT	30.7	Referential
22	IN TERMS OF	30.2	Discourse
23	TO BE A	29.9	Referential

Table 7 - BAWE Uniques

Rank	Bundle	Norm. freq	Function
14	THE MAJORITY OF	24.2	Referential
16	THE IDEA THAT	23.6	Referential
20	IT HAS BEEN	23.0	Referential
21	TO LOOK AT	23.0	Referential
23	SUCH AS THE	21.7	Discourse
28	IT IS IMPORTANT TO	19.9	Stance
29	IN THE CASE	19.2	Discourse
31	FOR EXAMPLE IN	18.6	Discourse
33	A VARIETY OF	18.0	Referential
34	THE STUDY OF	17.4	Referential

Tables 6 and 7 lists the ten most frequent bundles appearing only in one of the corpora. Appearing in the VESPA list does not necessarily mean that a construction is entirely absent from BAWE (or vice versa), however, but that it fails to meet the frequency or distribution criteria set for the present study. Identifying unique bundles, then, serves as an alternative approach to identifying severe over- or underuse of a bundle in VESPA. While the scope of the present study does not allow for an in-depth study of all these, there are certain patterns that emerge simply from comparing the two lists.

Firstly, many of the BAWE-unique bundles, such as *the majority of*; *the idea that*; *a variety of*; *the study of* are bundles that can be expected to appear as part of fairly complex noun phrases, indicating that the BAWE contributors are, to a greater extent than their VESPA counterparts, actively using constructions associated with structural compression, a feature central to academic writing (Biber & Gray 2010:6).

There are also certain immediate correspondences between the two lists. *A lot of* and *most of the*, both VESPA uniques, are plausible synonyms for the most frequent of the BAWE uniques, *the majority of*. Compare, for example, (13 and (14), where the two appear to be fairly interchangeable:

> (13) "However **the majority of the time** subjects did respond with a word in the same word class." BAWE 6174b

(14) “This is interesting because it can either mean that fine is used in a positive way **most of the time** .” VESPA UIO0236-LIN-01

The same seems to be the case for the BAWE-unique bundle “a variety of”, which can also be used in similar contexts to “a lot of”, as well as “ as seen in (15) and (16), and the VESPA-underused “a number of” (5.4).

(15) “Research has been carried out for many years **by a variety of different** researchers in connection with the differences between the speech of males and females and why these differences, if any, exist.” BAWE 6120e

(16) “Norwegian learners of English experiences **a lot of different** problems when writing in English” VESPA UIO0172-LIN-02

Two of these phrases, “the majority of” and *seems to be*, stand out as especially phraseologically interesting, the former being the most frequent of the BAWE-unique bundles, and the latter as a stance expression, which 5.3 and 5.4 show to be a challenging area for L2 users. These will be discussed in further detail below.

6 Discussion of individual bundles

6.1 On the other hand

The case of *on the other hand* is an interesting illustration of the challenges facing advanced learners. It is in no sense of the word a *wrong* construction, nor is it out of place in an academic setting. In fact, Liu (2013:33) lists *on the other hand* as one of the most common multi-word sequences in academic writing, the construction seems a recurring favourite with which to exemplify an academic discourse organising bundle (Biber & Conrad 2004:66; Chen & Baker 2010:38), and even the control corpus at hand confirms that *on the other hand* is a productive, useful phrase with which to organize a text in an academic setting. The problem, then, is not that the phrase in itself is not idiomatic, but that the massive overuse among the non-native writers – a log-likelihood test[4] of the two returns an LL value of 43.42 – suggests that not all of its 163 occurrences are in line with how a native speaker would use the phrase.

The underlying discoursal function, that of directing the flow of an argument, has been shown by Granger and Tyson (1996) to be an aspect of English with which L2 learners often struggle. By investigating the ICLE corpus, they found evidence of one-word connectors as a whole occurring with a similar frequency in both native and non-native writing, while certain connectors were both over- and underused. This pattern is remarkably similar to the findings of the present study, where the discourse organising function is equally frequent in both corpora but individual bundles are – sometimes drastically – overused.

Such overuse is indicative of what Hasselgren (1994) describes as a "lexical teddy bear"; a construction or word with which learners are comfortable, perhaps even overly so, and thus tend to use far more often than their native counterparts would. *On the other hand* corresponds well with a similarly used Norwegian term – *på den annen side*, and lends itself well to a nuanced, objective style of writing that illuminates both sides of a discussion, the virtues of which are extolled throughout Norwegian primary schooling, making it a safe, comfortable choice of construction for Norwegian L2 writers, prone to overuse.

[4] Log-likelihood calculations done with Paul Rayson's calculator. http://ucrel.lancs.ac.uk/llwizard.htm

It is worth remarking that some of this overuse can plausibly be attributed to the variance in assignments. A fair amount of the assignments in VESPA are comparative in nature, more or less prompting students to explicitly compare two or more alternatives, a context fairly conducive to the use of a contrastive phrase such as “on the other hand”. Such texts are less prominent in the BAWE sample. Still, these texts do not represent more than a fraction of the total occurrences, necessitating a more in-depth examination of the material.

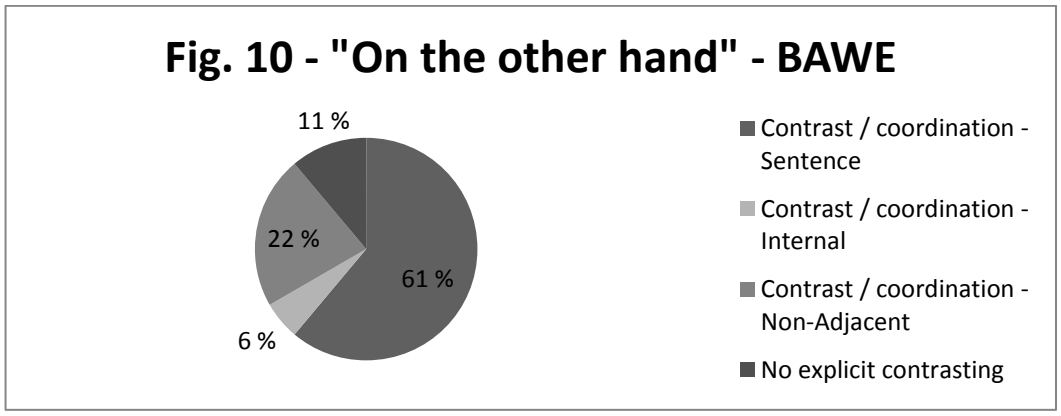
Investigating the 18 occurrences of the phrase in the BAWE corpus gives an indication of how the phrase is used by native speakers. Its main function in discourse is to introduce a comparison or contrast, often by way of juxtaposition, as (17) and (18) demonstrate.

(17) “Herbert concluded that compliments given by female speakers tend to have a personal base and use personal pronouns for example "I love your hat!" "You look great!" **On the other hand** compliments given by male speakers tend to be more impersonal: "Nice car!" or "Good job!" for example.” BAWE 6818b

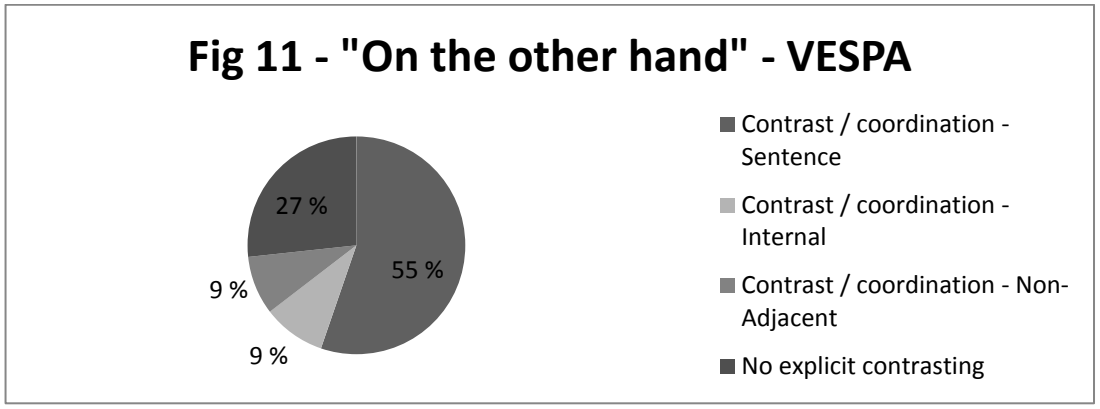
(18) “The social purpose of text B is interactional - a term normally used to refer to spoken discourse to describe the maintaining of social relations. Text A **on the other hand** is transactional, the purpose is to inform and request someone to do something.” BAWE 3127b

(19) “Where there are many types of verbs which are able to exhibit the pattern V of n, and it is largely the individual lexical items that carry the meaning, there are some patterns **on the other hand**, which themselves carry a certain degree of meaning.” BAWE 6061b

In both cases, the phrase is a cohesive device, both within and across sentences. As figure 10 shows, the majority of occurrences connect the preceding sentence to the current, primarily in a sentence-initial position (17). A select few occurrences provide contrast within the the same sentence (19) or refer to a non-adjacent sentence. Only two of the sentences have no obvious reference; they are used not for contrast, but simply to as a focusing aid, preparing the reader for a shift in subject matter.



Applying the same analysis to the VESPA sample reveals a similar ordering of functions, but in slightly different proportions, as seen in figure X. The most marked discrepancy between the two is the notable increase in non-contrastive constructions, with such constructions being more than twice as common in VESPA, accounting for one in four occurrences of the phrase, and thus also a significant portion of the overuse.



Turning to a more semantic approach, by examining the two concepts that are contrasted , a very distinct pattern emerges from the BAWE sample. In all the 16 contrastive occurrences, *on the other hand* serves to introduce a distinction between diametrically opposed concepts or qualities. Example 17 distinguishes male from female, and impersonal from personal, while (18) introduces the binary qualities of interactional and transactional, with the two characteristics mutually exclusive. This pattern of the preceding and succeeding concepts representing each side of a dichotomy is prevalent throughout the sample. In the majority of the occurrences, these binary variables are the only ones changed between the two declarative sentences, serving either to emphasise some fundamental difference or address an apparent contradiction. Example 17 again speaks of the differences between males and females, but continues to do so within the established topic of politeness, while example 18 stays within the realm of conversational interaction, despite shifting focus to another text type.

For the VESPA corpus, however, this strictly binary contrasting is not as dominant. VESPA contributors are as likely to use the phrase as a general topic change marker, introducing a concept only tangentially related to the preceding sentence, thereby changing not only certain variables, but the topic as a whole. The use seen in (20) is symptomatic, as the succeeding clause does not change a variable, it simply introduces a new argument, hence *on the other hand* does not mark an impending contrast or counter-argument, but simply works to inform the reader that the focus is about to shift from one aspect of a topic to another.

(20) “If it works you have substitution, and if we apply this test for this line, we find that it does. "We are the ones.." So far, so good. **On the other hand,** what is important to notice here is that we do not really know what the word is substituting for.” VESPA UIO0049-LIN-01

In the cases where VESPA users make such explicit contrastive comparisons, however, they do so by employing a unique tool, not found in the BAWE sample. Here, *on the other hand* represents the final half of a two-stage organising technique, following an earlier use of “on the one hand”. 17 out of the total occurrences are part of such pairings, as in (21), and there are in fact 6 additional uses of “on the one hand” where the second mention of “hand” is ellipted, as in (22).

(21) “**On the one hand** it is anaphoric reference to 'Africa Miracles' (19), **on the other hand** it can be argued that 'our' (20) also refers to 'good contacts in the country'” VESPA UIO0233-LIN-01

(22) “The text remains **on the one hand** explanatory, because it describes how the organization works, and interactive **on the other**, because it engages its audience directly.” VESPA UIO0062-LIN-01

This is a likely case of L1 transfer, as the construction “på den ene siden” and “på den andre siden” is a fairly common argument structure. Despite being entirely absent from BAWE, this construction is found in more general corpora, occurring in both the Corpus Of Contemporary American English and the British National Corpus, demonstrating again the L2 learners’ challenge of not only familiarising oneself with a phrase, but also the contexts in which it is idiomatic.

Three factors have thus been identified as contributing to the overuse of *on the other hand* in VESPA: the material gathered is skewed slightly toward explicitly comparative assignments; the non-contrastive use which remains infrequent in native writing is used more frequently by the VESPA contributors; and finally that the phrase is given a second function, that of topic introduction rather than elaboration. There is, however, a compelling alternative

explanation: Comparing the entirety of BAWE to a subsection of the FLOB consisting of published academic texts, Chen and Baker (2010:66) found that *on the other hand* is in fact one of the most frequent bundles in the FLOB sample, almost five times more frequent than in BAWE, suggesting that the contributors to the VESPA sample's "overuse" of the phrase may in fact be *more* in line with the conventions of academic writing than its BAWE counterpart.

6.2 The majority of

The majority of is the most frequent of BAWE's "unique" bundles, and 14th overall, with 24 occurrences per 100,000 words. Although not frequent enough to qualify as a lexical bundle, the phrase does occur 20 times[5] – 5.4 times per 100,000 words – in VESPA, implying an significant underuse of the phrase in VESPA, with a log-likelihood value of (negative) 32.01 Although the sample is limited, it can still be used to trace the contexts in which the Norwegian writers do in fact use the phrase.

While the basic function of the phrase is the same throughout, that of a quantifying determiner in a noun phrase, the other constituents of the noun phrase contribute greatly to the meaning. *Postmodified* phrases -make the reference more specific, specifying that the head of the noun phrase is the majority only of a certain sample of a clearly defined population, as in example (23 – postmodifier italicized),. The addition of a second determiner in the form of a *definite* article (italicized) has a similar specifying effect, in the form of an anaphoric reference (24), indicating that the sample discussed has been more clearly delineated at a previous point in the text. *Blank* phrases (25) contain only the phrase and the head noun, and are as such more general, leaving it up to the reader to infer the details from the context. Applying the same analysis to BAWE adds a fourth construction, where a *premodifier* (italicized) serves to specify the reference in a manner similar to that of the postmodification pattern (26)

(23) "**The majority of** marked themes *in Annan's lecture* denote settings realized by adverbials." VESPA UIO0002-LIN-01

[5] A total of 36 occurences are found in VESPA, but 16 of these are in reference to a text segment analysed in one of the assignments given, demonstrating the value of the tagset applied to exclude sequences not produced by the writer.

(24) "**The majority of** *the* referents copious modifies in these types of texts have to do with liquid [...]" VESPA UIO0025-LIN-03

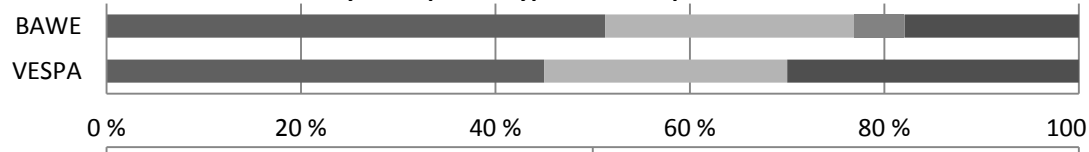
(25) "There are, in all, nine out of twelve instances of coordination - ing clauses in the corpus that express notions of immediacy, and in eight of them **the majority of** *translators* use *og*" VESPA UIO0001-LIN-05

(26) "[...] in order to establish whether the data presented in this study is, indeed, representative of **the majority** of *healthy, premature* infants." BAWE 6206b.xml

The distribution of these functions in both corpora is shown in fig. 11 below , and shows that with the notable exception of premodified phrases, non-native users seem fairly comfortable with the contexts in which *the majority of* can be used. A relative overuse of the postmodified construction and a similar underuse of the unmodified function is evident, indicating that the VESPA contributors tend to prefer explicit over contextually implied references. Investigation of these examples do however reveal that there are other means of specification than noun phrase modification, as the anaphora in 24 above illustrates. A new, more pragmatic approach is therefore necessary in order to ascertain whether or not this hypothesis holds true.

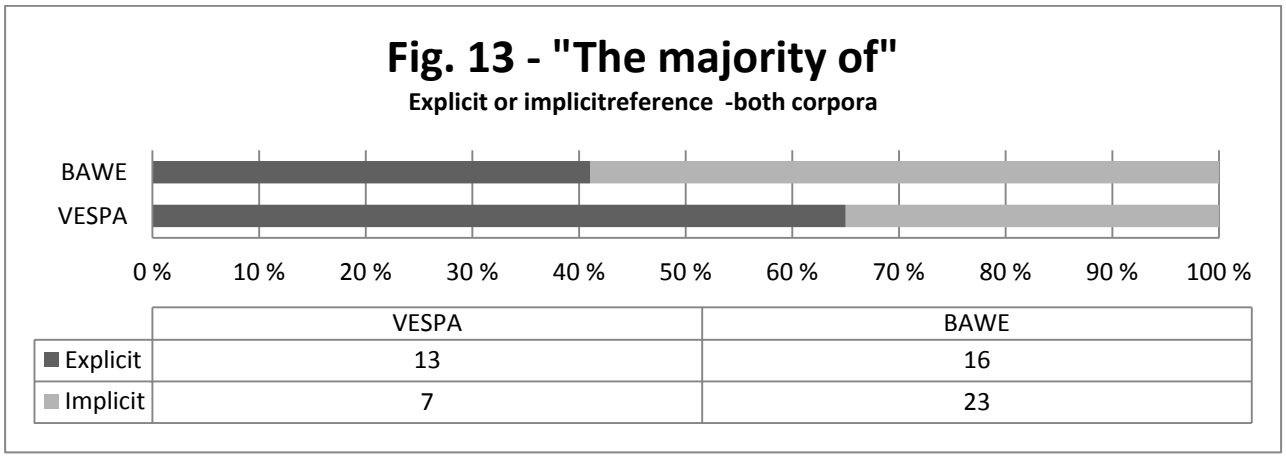
Fig. 12 - "The majority of"

By noun phrase type - both corpora



	VESPA	BAWE
Blank	9	20
Definite	5	10
Premodified	0	2
Postmodified	6	7

To this end, occurrences are sorted according to a relatively simple functional criterion: is the reference given explicitly in the sentence, or implied by the context? The results are seen in figure 13, and show a significant divergence between the two samples.



	VESPA	BAWE
Explicit	13	16
Implicit	7	23

While the implicit function is the dominant one in BAWE, accounting for more than half the sample, the opposite is the case in VESPA, where two out of three occurrences state their reference explicitly. Typical of the VESPA use is reference to a majority within a sharply delineated population already referred to or in some other way presented by the writer, either by expanding the noun phrase, as in (27), or by reference elsewhere in the sentence (28). Example 29 seems almost bold by comparison, simply stating that their claim is true for "the majority of cases", sacrificing accuracy for efficiency, expecting the reader to infer the scope of the sample discussed.

(27) "[...]both in BrE and AmE texts the assumed BrE punctuation practice has been observed in **the majority of** *cases used for the analysis* which in the case of AmE is a deviation from norm stated in theoretical sources." VESPA UIO0004-LIN-02

(28) "[...] *particularly for SWICLE* where **the majority of** the essays were written under exam conditions with a time limit." VESPA UIO0083-LIN02

(29) "This is as a result of the fact that in **the majority of** cases, these two tenses are characterised solely by inflectional endings and therefore although they have to fit logically into the sentence, there is only one word compared to several when constructing the future tense." BAWE6120c

A proficient writer of English is not automatically a proficient writer of academic English ,however, and, assuming that accuracy in description is a desirable quality, it could be argued that the VESPA use is in fact *more* academic than the BAWE use, since it communicates in a more precise manner, less open to misunderstandings and ambiguity. Biber and Gray's 2010 study comparing spoken and academic language concludes otherwise, however,

demonstrating that an inexplicit writing style is in fact a key component of expert academic texts. The lack of explicitness, they claim, "causes few, if any problems, because the expert reader anticipates the expected readings that will occur in this context" (2010:18) , while the compact, efficient style that results from such inexplicit constructions enables researchers to "keep up with the volume of information produced by scientific researchers"(2010:19).

In dictionaries, the more general use of the term is conspicously absent. "The MacMillan online dictionary, for example, defines "majority" as "most of the people or things in a group[6]", which is more in line with the more restricted VESPA use, suggesting that this could also be a case of semantic bleaching, a "change by which the meaning of a word becomes increasingly unspecific" (Matthews 2007:43). Should this be the case, the VESPA underuse of the phrase is not necessarily unidiomatic, but more a case of a recent language development that has yet to spread to L2 users.

This assymetry is a strong indication that Norwegian L2 writers are more cautious than necessary in their use of the "majority" term, commonly reserving it for contexts in which a referent constitutes an empirically provable majority. A possible reason for this could be a transfer effect from the relatively formal use of the Norwegian cognate "majoriteten", which, somewhat interchangably with its Germanic counterpart "flertallet" is used primarily to refer to clearly defined majorities, often of a political nature. More abstract majorities, such as "the majority of children" - of which there are four occurrences in BAWE - are more likely to be referred to by more general terms such as "de fleste" in Norwegian, a possible explanation for the overuse of similar, less formal quantifying determiners discussed in 5.5. Whatever the underlying causes, however, the corpus data clearly indicate that this reluctance to use *the majority of* in inexplicit contexts combined with an overreliance on less formal quantifiers are contributo significantly to the under-use of *the majority of* in VESPA, which in turn contributes to a style of writing less suited for academic purposes.

[6] Available online at http://www.macmillandictionary.com/dictionary/british/majority

6.3 *Seems to be*

As the most frequently used bundle unique to VESPA and a representative of the seemingly troublesome stance expression category, the apparent overuse of *seems to be* is one that warrants a further investigation. The basic function of the phrase is one of a copular verb expressing epistemic modality, providing an alternative to the more absolute "is" in settings in which the writer wishes to express a degree of uncertainty, insecurity or other factual assessment. As with *on the other hand*,the phrase is present in BAWE, but far less frequent, with a log-likelihood test result of 13.85 indicating a significant VESPA overuse.

Turning to the the relatively limited set of occurrences in the BAWE corpus – a total of 12 - in order to investigate the manner in which native users employ the phrase allows for the identification of two distinct functions of the phrase: the expression of uncertainty and insecurity.

The expression of *uncertainty* is largely owed to the abstract, intangible or inconclusive nature of the subject matter. In example (30) the matter at hand is that of ascertaining whether or not the role of decoding is more or less important in current reading theory than it has previously been, and the modal *seem* is employed partly because the writer does not claim to have read *all* current reading theory and is thus unable to make an absolute claim, but also because "important" is a subjective, intangible term that cannot necessarily be quantified and summarised across all relevant sources. Similarly, in (31), the writer is making an assumption rather than an empirically founded conclusion, and opts for a modal expression to emphasise the tentative nature of the claim. The modal expression then, is a necessity.

> (30) "The role of decoding therefore **seems to be** more important in current reading theory" BAWE 6174e"
>
> (31) "The high number of misspelled words in this essay **seems to be** due to slips rather than actual errors". VESPA UIO0036-LIN-02

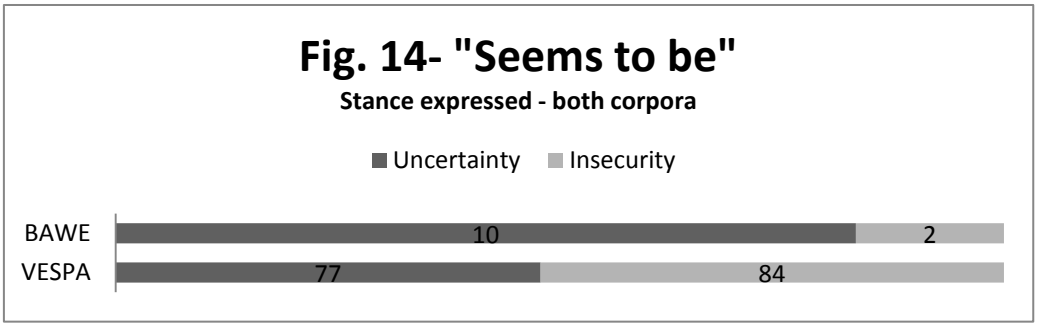
The other stance expressed by the phrase is quite different, expressing not uncertainty, but *insecurity*. In these cases, the modality serves to soften an argument more than it expresses any uncertainty with regard to the truth of a claim, as in (32), where the writer has in fact accounted for the empirical truth of the matter, having a clearly defined sample in which a translation has been identified as "most typical", yet continues to use the modal expression, most likely as a way of distancing him- or herself should this analysis prove to be wrong. (33)

uses modality to a similar effect, as the frequency of "thus" in the two data sets referred to is in fact accounted for in the preceding text. The modal expression is no longer a necessity, but a matter of retaining a humble, non-insistent style.

> (32) "The most typical word used in the translations **seems to be** may/might followed by one or more verbs in the present tense or the present continuous, which was used in thirty percent of the translations." UIO0033-LIN-01

> (33) "Also *thus* **seems to be** more common in texts translated from Norwegian into English than in English source texts and here it particularly applies to non-fiction." VESPA UIO0089-LIN-01

Sorting all occurrences according to the stance expressed provides a vivid illustration(Fig. 14) of how the use of *seems to be* varies drastically between the two corpora:



Despite the relatively small BAWE sample, these results are striking, indicating that the non-native contributors are more liable to use *seems to be* as an expression of insecurity rather than uncertainty, while this function only accounts for one in six uses of the same phrase in the native sample.

Several variations of this seemingly superfluous epistemic modality use can be found among the 84 "Insecure" uses. Most frequent are constructions similar to (34), where fairly straightforward facts are given with a measure of insecurity, where the writer does not appear to be qualifying the statement in any manner, but simply admitting to not being entirely sure that the claim is "correct". Other errors are more stylistic, as with (35), where the writer is referring to a previous hypothesis given in the form of a modal construction more reminiscent of an initial observation than a hypothesis, or (36), where a prepositional phrase is somewhat paradoxically said to "seem clearly optional".

> (34) "*It* **seems to be** an anaphoric reference which refers to "travelling in Morocco" VESPA UIO0230-LIN-01

(35) “This does not fit my hypothesis that totally is a word that **seems to be** "overused" by Norwegian learners of English.” UIO0029-LIN-01

(36) “This can be seen in the examples ENG1 and ENG2 above where the prepositional phrase with you in ENG2 seems to be clearly optional rather than obligatory.” UIO0027-LIN-01

This seeming lack of confidence among the L2 contributors is especially noteworthy when considering that the two samples are both apprentice writers of academic English, and should be, at least in theory, equally comfortable with the conventions of the academic genre. Although this can likely be contributed at least partly to cultural differences, it also suggests that the fact that the VESPA contributors are expressing themselves in a foreign language as well as within a set of unfamiliar, largely unwritten, conventions, could lead to a lack of confidence and resulting hesitance in stating their as fact, although the judgments they make are as academically founded as the BAWE contributors’.

This variety of alternative uses of the phrase and the VESPA contributors’ affinity for more tentative, cautious constructions, be they cultural or otherwise, seems to go far in explaining the considerable overuse of the stance expression *seems to be* compared to BAWE, and seems to hold true even if extending the search to *seem* and its inflected forms in isolation, with 687 hits in VESPA and 173 in BAWE, suggesting a continued overuse, with a log-likelihood value of 45.64.This harmonizes well both with the findings of chapters 5.2 and 6.3 of the present study, suggesting that Norwegian advanced learners, although certainly capable of making their own attitudes known through stance expressions, have some way to go before mastering the elusive art of native-like selection.

7 Conclusion

7.1 Summary of findings

In comparing limited samples of the VESPA and BAWE corpora, the present study has been able to identify several characteristics of how Norwegian students of English linguistics and literature use English lexical bundles in academic texts. These findings are largely in line with earlier relevant studies, such as Chen and Baker 2010, Ädel and Erman 2012 and Biber et al. 2004, although it is most so with the last, where the referential category was seen to be equally dominant. This demonstrates not only the merit of the lexical bundle approach , as it can be seen to produce fairly similar results when investigating similar fields, but also that there definitely seems to exist a notion of an academic norm.

Returning to the initial research questions, is is evident that the Norwegian contributors are in fact using lexical bundles, and that a considerable amount of these are the same bundles preferred by native writers. The functions served by these bundles are strikingly similar in both corpora, with the majority of bundles performing a referential function, while a comparatively small, yet crucial set of bundles contribute to the organisation of the text and the expression of the writers' attitudes and judgments.

Although the lexical bundles employed in VESPA are similar in function to those used by native writers, the Norwegian learners have demonstrated a tendency of overreliance on certain bundles, a pattern that is especially in their use of stance expressions. With a considerably smaller set of lexical bundles at their disposal, the manner in which VESPA contributors' communicate their attitudes and uncertainties is frequently diverges from that of their native writer counterparts. This manifests itself in an overly cautious style, replete with qualified statements, hedgings and explicitly stated information inferable from the context, as illustrated by marked uses both of *the majority of* (6.2) and *seems to be* (6.3). Idiomatic discourse organisation, however, is characterised by the exact opposite, as bundles are applied in too general a manner, as seen in the case of *on the other hand* (6.1)

In terms of adherence to academic convention, the VESPA contributors are again underperforming slightly when compared to the BAWE contributors. Features central to academic writing, such as structural compression and an inexplicit writing style are less

common, and where alternative, less formal constructions are available, VESPA contributors show a tendency of preferring these over more characteristically academic bundles, exemplified by an underrepresentation of passive constructions and certain crucial academic bundles, as illustrated in 5.3 and 5.4.

7.2 Limitations and suggestions for further research

The present study is subject to certain limitations. First and foremost, the sample under investigation represents a fairly narrow field of student writing, and the applicability of the conclusions derived from it are thus not necessarily valid for other demographics, disciplines or L1s. Secondly, the classification scheme of the lexical bundle framework is still very much a preliminary framework, and would benefit greatly from the implementation of more rigid criteria for category assignments.

Nevertheless, it is the opinion of the present author that a comparative study of the BAWE and VESPA corpora has definite merit, with several salient insights gained from the present study. Similar analyses of the other L1 subcorpora of VESPA are encouraged, as they could, in addition to the obvious applications for both teachers and learners of the relevant languages, prove valuable for the study of EAP as whole, as their results combined would allow for the identification of challenges common to learners of academic English regardless of L1 background.

As seen in figure 2, almost a third of the bundles retrieved were omitted from the present study, as they did more to illustrative of the subject matter of the corpus texts that they did to illuminate the idiosyncrasies of non-native EAP use. A fair shared of these bundle come in the shape of a lexical word flanked by function words, however, and the identification of the *lexical frames* that these function words create invites to a wealth of potentially interesting investigations, as recently demonstrated by Gray and Biber (2013.

7.3 Pedagogical implications

The present study has highlighted two specific areas of academic writing that prove challenging to Norwegian learners of English: Stance expressions and discourse organisers. For both categories, the fairly limited repertoire of relevant constructions restrict the options available when writing, and lead to severe overuse of certain terms, often leading to unidiomatic uses, as seen with *on the other hand* and *seems to be*, while other constructions more suitable to academic discourse are underused, such as *due to the, in order to* and *a number of*. Equipping students with a wider repertoire of such constructions, and teaching them the various contexts in which they are and are not suitable should be a priority for students that appear to be approaching a near-native level of competence.

Perhaps as important as explicit vocabulary building, however, is the teaching of basic academic conventions. Not only the basics of the inexplicit, contextualised writing style rich in structural compression and passive constructions that has been shown to characterise academia, but also how to accurately express oneself in academic terms, how to be general without being ambiguous, when to be cautious, and perhaps most importantly, when not to be. It should be the objective of teachers of academic English to equip learners with the tools necessary to know when it is time to eschew “the fact that the majority seems to be”, for a simple fact: “the majority is”.

Bibliography

Ädel, Annelie and Britt Erman. "Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach." *English for Specific Purposes* 31.1 (2012): 81-92.

Ädel, Annelie. *Metadiscourse in L1 and L2 English*. Amsterdam: Benjamins, 2006.

Alsop, Sian and Hilary Nesi. "Issues in the development of the British Academic Written English (BAWE) corpus." *Corpora* 4.1 (2008): 71-83.

Altenberg, Bengt. "On the Phraseology of Spoken English: The Evidence of Recurrent Word-combinations." Cowie, Anthony P. *Phraseology. Theory, Analysis and Applications*. Oxford: Oxford University Press, 1998. 101-122.

Biber, Douglas. "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing." *International Journal of Corpus Linguistics* 14.3 (2009): 275-311.

Biber, Douglas and Bethany Gray. "Challenging stereotypes about academic writing: Complexity, elaboration, explicitness." *Journal of English for Academic Purposes* 9.1 (2010): 2-20.

Biber, Douglas and Federica Barbieri. "Lexical bundles in university spoken and written registers." *English for Specific Purposes* 26.3 (2007): 263-286.

Biber, Douglas, et al. *Longman Grammar of Spoken and Written English.* London: Longman, 1999.

Biber, Douglas, Susan Conrad and Viviana Cortes. "If you look at …: Lexical bundles in university teaching and textbooks." *Applied Linguistics* 25.3 (2004): 371–405.

Biber, Douglas, Susan Conrad and Viviana Cortes. "Lexical bundles in speech and writing: an initial Taxonomy." Wilson, Andrew, Paul Rayson and McEnery Anthony (eds.). *Corpus Linguistics by the Lune - A Festschrift for Geoffrey Leech*. Frankfurt am Main: Peter Lang verlag, 2003. 71-93.

Chen, Yu-Hua and Paul Baker. "Lexical Bundles in L1 and L2 Academic Writing." *Language Learning & Technology* 14.2 (2010): 30-49.

Conrad, Susan and Douglas Biber. "The Frequency and Use of Lexical Bundles in Conversation and Academic Prose." *Lexicographia* 20 (2004): 56-71.

Cowie, Anthony P. *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 1998.

De Cock, Sylvie. "Preferred sequences of words in NS and NNS speech." *Belgian Journal of English Language and Literatures (BELL)* (2004): 225-246.

Ebeling, Jarle, Signe Oksefjell Ebeling and Hilde Hasselgård. "Using recurrent word-combinations to explore cross-linguistic differences." Aijmer, Karin and Bengt Altenberg (eds.). *Advances in Corpus-based Contrastive Linguistics: Studies in honour of Stig Johansson* . . Amsterdam: John Benjamins Press, 2013. 177-200.

Ebeling, Signe and Alois Heuboeck. "Encoding document information in a corpus of student writing: the British Academic Written English corpus." *Corpora* 2.2 (2007): 241-256.

Ellis, Nick C, Rita Simpson-Vlach and Carson Maynard. "Forumulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics and TESOL." *TESOL Quarterly* 42.3 (2008): 375-396.

Fiedler, Sabine. *English Phraseology. A Coursebook.* Tübingen: Gunter Narr Verlag, 2007.

Flowerdew, John. "Signalling nouns in discourse." *English for Specific Purposes* 22.4 (2003): 329-346.

Granger, Sylviane and Magali Paquot. "Disentangling the phraseological web." Granger, Sylviane and Francoise (eds) Meunier. *Phraseology. An Interdisciplinary Perspective*. Amsterdam: John Benjamins Publishing Company, 2008. 27-50.

Granger, Sylviane and Stephanie Tyson. "Connector usage in the English essay writing of native and non-native EFL speakers of English." *World Englishes* 15.1 (1996): 17-27.

Granger, Sylviane. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora." Aijmer, Karin (ed.). *Languages in Contrast. Text-based cross-linguistic studies*. Lund: Lund University Press, 1996. 37-51.

Granger, Sylviane. "Computer Learner Corpus Research: Current Status and Future Prospects." Connor, Ulla and Thomas A. (eds.) Upton. *Applied Corpus Linguistics - A Multidimensional Perspective*. Amsterdam: Benjamins, 2004. 123-146.

Granger, Sylviane. "From Phraseology to Pedagogy: challenges and prospects." Herbst, Thomas, Susen Faulhaber and Peter (eds.) Uhrig. *The Phraseological View of Language*. Berling: De Gruyter Mouton, 2011. 123-146.

Granger, Sylviane. "How to Use foreign and Second Language Learner Corpora." Mackey, Alison and Susan M (eds.) Gass. *Research Methods in Second Language Acquisition*. Chichester: Wiley Blackwell, 2012. 7-29.

Granger, Sylviane. "Learner Corpora." Chapelle, Carol A. *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell Publishing, 2013.

Gray, Bethany and Douglas Biber. "Lexical frames in academic prose and conversation." *International Journal of Corpus Linguistics* 18.1 (2013): 109-135.

Haislund, Niels. "Otto Jespersen." *Englische Studien* 75 (1943): 273-283.

Hasselgård, Hilde and Stig Johansson. "Learner corpora and contrastive interlanguage analysis." Meunier F., De Cock S., Gilquin G. and Paquot M. (eds). *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: Benjamins, 2011. 33-62.

Hasselgreen, Angela. "Lexical Teddy Bears and Advanced learners: a study into the ways Norwegian students cope with vocabulary." *International Journal of Applied Linguistics* 4.2 (1994): 237-260.

Howarth, Peter. "Phraseology and Second Language Proficiency." *Applied Linguistics* 19.1 (1998): 24-44.

Hughes, Rebecca. *English in speech and writing: investigating language and literature*. London: Routledge, 1996.

Hyland, Ken. "Applying a gloss: exemplifying and reformulating in academic discourse." *Applied Linguistics* 28.2 (2007): 266-285.

Hyland, Ken. "As can be seen: Lexical bundles and disciplinary variation." *English for Specific Purposes* 27.1 (2008): 4-21.

Johansson, Stig. "On the Role of Corpora in Cross-Linguistic Research." Johansson, Stig and Signe Oksefjell. *Corpora and Cross-Linguistic Research, Theory, Method and Case Studies*. Amsterdam: Rodopi, 1999. 3-24.

Johansson, Stig. "Some aspects of the development of corpus linguistics in the 1970s and 1980s." Lüdelig, Anke and Merja Kytö. *Corpus Linguistics - An International Handbook*. Berlin / New York: Walter de Gruyter, 2008. 33-3.

Keen, J. "Sentence-combining and redrafting processes in the writing of secondary school students in the UK." *Linguistics and Education* 15 (2004): 81-97.

Kucera, Henry and Francis Nelson. *Computational Analysis of Present-day American English*. Providence: Brown University Press, 1967.

Leech, Geoffrey. "Corpora and theories of linguistic performance." Svartvik, Jan (ed.). *Directions in corpus linguistics: proceedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter, 1992. 105-122.

Leech, Geoffrey. "The state of the art in corpus linguistics." Aijmer, Karin and Bengt (eds.) Altenberg. *English Corpus Linguistics: Studies in honour of Jan Svartvik*. London: London, 1991. 8-29.

Li, Li-Juan and Guang-Chun Ge. "Genre analysis: Structural and linguistic evolution of the English-medium medical research article(1985-2004)." *English for Specific Purposes* 28.2 (2009): 93-104.

Lin, Ling and Stephen Evans. "Structural patterns in empirical research articles: A cross-disciplinary study." *English for Specific Purposes* 31.2 (2012): 150-160.

Liu, Dilin. "The most frequently-used multi-word constructions in academic written English: A multi-corpus study." *English for Specific Purposes* 31.1 (2012): 25-35.

Matthews, P.H. *Oxford Concise Dictionary of Linguistics*. Oxford: Oxford University Press, 2007.

McEnery, Anthony M and Andrew Hardie. *Corpus Linguistics. Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012.

McEnery, Anthony M and Andrew Wilson. *Corpus Linguistics (second edition)*. Edinburgh: Edinburgh University Press, 2001.

Nesi, Hilary et al. *An Investigation of Genres of Assessed Writing in British Higher Education: Full research report. ESRC End of Award Report, RES-000-23-0800*. Swindon: ESRC, 2008.

Paquot, Magali. "Exemplification in learner writing: a cross-linguistic perspective ." Granger, Sylviane and Fanny (eds.) Meunier. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: Benjamins, 2008. 101-119.

Paquot, Magali, Hilde Hasselgård and Signe Oksefjell Ebeling. "Writer/reader visibility in learner wrtiting across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora." Granger, Sylviane, Gaetanelle Gilquin and Francoise (eds) Meuiner. *Twenty Years of Learner Corpus Research: Looking back, Moving avhead. Corpora and Language in Use - Proceedings 1*. Louvain-la-Neuve: Presses universitaires de Louvain, 2013. 377-387.

Pawley, Andrew and Frances Hodgetts Syder. "Two puzzles for linguistic theory: nativelike selection and nativelike fluency." Richards, S.C and R.W. (eds.) Schmidt. *Language and Communication*. London / New York: Longman, 1983. 191-226.

Pendar, Nick and Carol A. Chapelle. "Investigating the Promise of Learner Corpora: Methodological Issues." *CALICO Journal* 25.2 (2008): 189-206.

Scott, Mike. "Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs." Ghadessy, Mohsen, Alex Henry and Robert L (eds) Roseberry. *Small Corpus Studies: Theory and Practice*. Amsterdam: Benjamins, 2001. 47-67.

Sinclair, John. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

Sinclair, John. "EAGLES Preliminary Recommendations on Corpus Typology." 1996. Web page. 9 December 2013. <http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html>.

Sinclair, John. *Trust the Text: Language, Corpus and Discourse*. London: Routledge, 2004.

Sinclair, John. "Corpus and Text - Basic Principles." Wynne, Martin (ed). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 2005. 1-16.

Staples, Shelley, et al. "Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section." *Journal of English for Academic Purposes* 12.3 (2013): 214-225.

Stubbs, Michael and Isabel Barth. "Using Recurrent Phrases as Text-type Discriminators." *Functions of Language* 10.1 (2003): 65-108.

Stubbs, Michael. "Quantitative data on multi-word sequences in English. The case of the word "World"." Hoey, Michael, et al. *Text, Discourse and Corpora. Theory and Analysis*. London: Continuum, 2007. 163-189.

Swales, John M. "When there is no perfect text: Approaches to the EAP practitionar's dilemma." *Journal of English for Academic Purposes* 8.1 (2009): 5-13.

Tognini-Bonelli, Elena. *Corpus linguistics at work*. Amsterdam: John Benjamins Press, 2001.

Tribble, Christopher and Michael Scott. *Textual Patterns: Key words and Corpus Analysis in Language Education*. Amsterdam: Benjamins, 2006.

Tribble, Christopher. "Revisiting apprentice texts: Using lexical bundles to investigate expert and apprentice performances in academic writing." Meunier, Fanny, et al. *A Taste for Corpora - In honour of Sylviane Granger*. Amsterdam: Benjamins, 2011. 85-108.

Wright, Laura J. "Writing science and objectification: Selecting, organizing, and decontextualing knowledge." *Linguistics and Education* 19.3 (2008): 265-293.

Wulff, Stefanie and Ute Römer. "Becoming a proficient academic writer: shifting lexical preferences in the use of the progressive." *Corpora* 4.2 (2009).